Chromatin conformation and chromosomal organization during the activation of T-cells

Antonis Klonizakis, M.Sc programme “Molecular biology and Biomedicine” thesis

Supervisors :

Christoforos Nikolaou , Assistant Professor of Bioinformatics, University of Crete, Department of Biology

Charalampos Spilianakis, Associate Professor of Molecular Biology, University of Crete, Department of Biology

Table of contents

Στον μέντορα και καθηγητή μου, Χριστόφορο.

Για το ότι δέχθηκε ένα παιδαρέλι στο εργαστήριο του πριν από τρία χρόνια.
Για το ότι μου έμαθε το πως θα έπρεπε να είναι η έρευνα και το πώς σκέφτεται ένας επιστήμονας.
Για το ότι μου έμαθε το πως μοιάζει ένας μέντορας.

Στον μέντορα και καθηγητή μου κ. Σπηλιανάκη.

Για την εμπιστοσύνη και την βοήθειά του.
Για το ότι με έμαθε να ενθουσιάζομαι με αυτό που κάνω.

Στους ανθρώπους και των δύο εργαστηρίων που μου έμαθαν το πως θα έπρεπε να μοιάζει κάθε εργασιακός χώρος.

Στον Τόμας, που μοιράστηκε μαζί μου τα δεδομένα της πολυετής δουλειάς του.

Στην Λυδία και στον Αιμίλιο που έκαναν την κάθε μέρα στο εργαστήριο να μετράει.
Στους δυο τους και για τις μέρες εκτός του εργαστηρίου.

Στον Στέλιο, στον Κωνσταντίνο, στη Μυρσίνη, στη Μαρία, στη Λαμπρίνα, στη Μανουέλα, στην Ντέπη, στον Γιώργο, στη Λητώ και στον Νίκο.

Στα νέα μέλη του CGG, την Ίλια, την Ελένη τη Σοφία και την άλλη Σοφία, που διατηρούν το οικογενειακό κλίμα του εργαστηρίου.

Στους συμφοιτητές μου, που πέρασαν μαζί μου δύο χρόνια εντός και εκτός της άιθουσας Α.

Στους φίλους μου στο Ηράκλειο που μου έμαθαν όλα τα υπόλοιπα τα τελευταιά έξι χρόνια.

Στον Άγγελο, στον Λεωνίδα, στη Στέλλα, στον Πέτρο και στην Κατερίνα που μένουν ακόμα εδώ.

Στη Σταυρούλα, στην Ελένη, στον Σίμο, στον Βαγγέλη και στον Γιώργο που μένουν μακριά από εδώ.

Στη Στεφανία που μένει και αυτή μακριά από εδώ.

Περίληψη

Η γονιδιακή ρύθμιση και έκφραση αποτελούν κρίσιμες διαδικασίες που ελέγχουν όλες τις πτυχές της ζωής των ευκαρυωτών. Η τρισδιάστατη δομή του γονιδιώματος αποτελεί πλέον ένα επιπρόσθετο επίπεδο ρύθμισης της έκφρασης γονιδίων, μέσω μηχανισμών που δεν είναι ακόμα καλά μελετημένοι. Στην παρούσα εργασία, μελετήθηκε ο ρόλος της πρωτεΐνης SATB1 σε κύτταρα ποντικού. Το γονίδιο *Satb1* εκφράζεται σε συγκεκριμένους κυτταρικούς τύπους, όπως τα διπλά θετικά Τ-κύτταρα του θύμου αδένα. Η πρωτεΐνη αυτή έχει συσχετισθεί με την δημιουργία “λουπών’’ και άλλων τρισδιάστατων χρωματινικών δομών, οι οποίες ρυθμίζουν την έκφραση συγκεκριμένων γονιδίων. Χρησιμοποιώντας μια διαγονιδιακή σειρά ποντικών, από την οποία απουσιάζει το γονίδιο της *Satb1* στα κύτταρα του θύμου αδένα, δείχνουμε πώς τα ζώα αυτά αναπτύσσουν αυτοανοσία. Με τη χρήση πειραμάτων αλληλούχισης νέας γενιάς, γίνεται μια ενδελεχής μελέτη του μοριακού προφίλ των θυμοκυττάρων σε ποντίκια που έχουν απωλέσει το γονίδιο. Παρατηρούμε εκτεταμένες αλλαγές στην έκφραση γονιδίων, στην προσβασιμότητα της χρωματινής, στα επίπεδα μεθυλίωσης του DNA, αλλά και στις τρισδιάστατες αλληλεπιδράσεις μεταξύ πολλών περιοχών του γονιδιώματος. Προτείνουμε ένα μοντέλο, σύμφωνα με το οποίο, η απώλεια του γονιδίου της *Satb1* οδηγεί σε διακοπή της διαφοροποίησης των Τ-λεμφοκυττάρων και στο σταδιακό εκφυλισμό των επιθηλιακών κυττάρων του θύμου εξαιτίας της δυσλειτουργικής κυτταρικής επικοινωνίας μεταξύ των Τ-κυττάρων και των επιθηλιακών κυττάρων αυτών. Τα παραπάνω φαινόμενα εξηγούν μερικώς την ανάπτυξη αυτοανοσίας σε περίπτωση απώλειας του γονιδίου.

Abstract

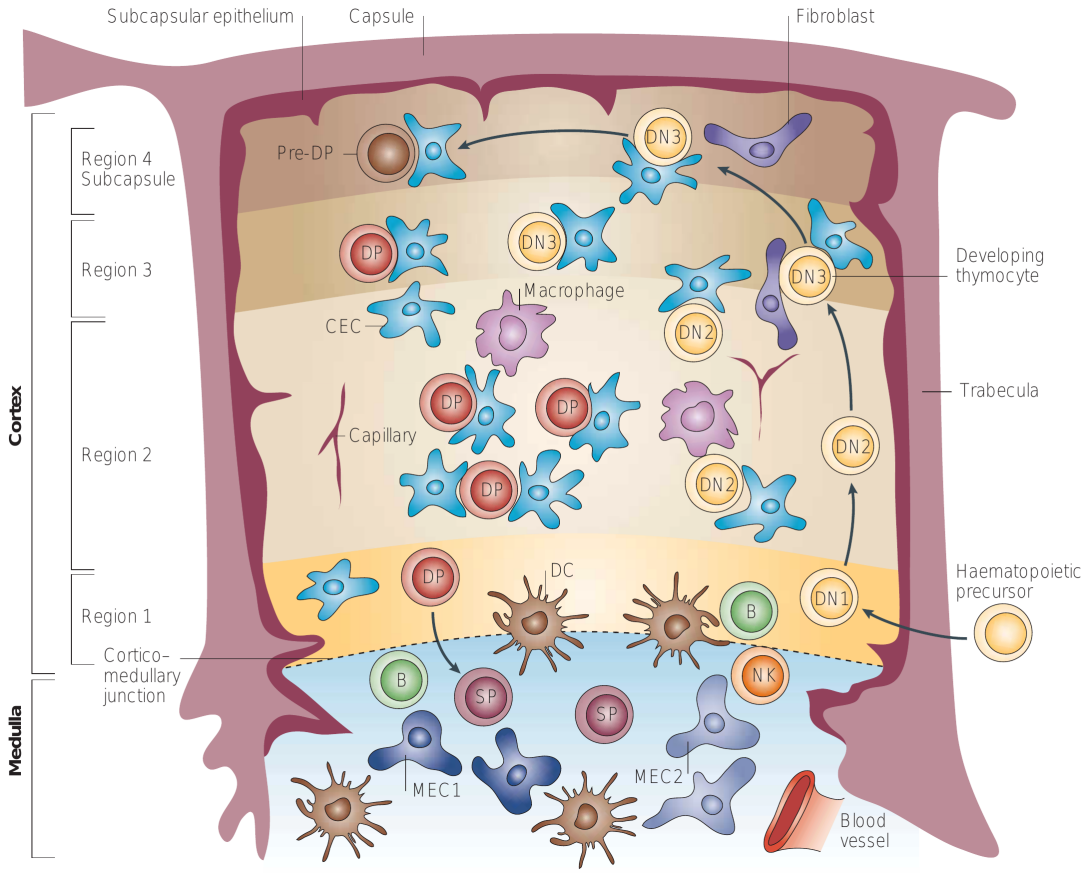
Gene expression and regulation represent crucial processes that control every aspect of eukaryotic life. 3D genome organization has emerged as an additional layer of transcriptional control, through mechanisms that are not well perceived. In this work, SATB1 was studied in mouse cells. *Satb1* is expressed in specific cells, such as double positive T-cells of the thymus. SATB1 has been shown to create loops and other three-dimensional structures that ultimately lead to the regulation of gene expression. Using a conditional knockout mouse line, in which the gene is deleted from double positive T-cells, we show autoimmunity development. Using next generation sequencing experiments, we describe in detail the molecular profile of thymocytes that have abolished *Satb1*. Widespread changes were observed in the expression levels of genes, accessibility of chromatin, levels of DNA methylation but also in the three dimensional contacts between multiple genomic loci. We propose a model, according to which, the loss of *Satb1* in thymocytes leads to the developmental arrest of double positive T-cells and to the degeneration of the thymus due to malfunctioning cell to cell communication events between T-cells and thymus epithelial cells. The above findings partially explain the development of autoimmunity following the loss of *Satb1* in thymocytes.

Introduction

The adaptive immune system is a crucial subtype of the immune system of vertebrates. Adaptive immunity allows highly specific immune responses versus pathogens that are hard to deal with by other clearance mechanisms, like those of innate immunity. Although an adaptive response takes a lot of time to develop, since it relies on rather sophisticated mechanisms, it is extremely crucial for the survival of each individual : An infant born with a defective adaptive immune system is doomed to die soon after its birth (Bruce Alberts et al., 2002).

The thymus is an important organ of the adaptive immune system. The proper development of T-lymphocytes (T-cells from now on), occurs in this bilobed organ. T-cells are important components of the adaptive immune system carrying out a wide variety of functions. Some T-cell subsets stimulate B-cells for antibody production, others restrict immune responses and other mediate cell death of virus-infected cells of the host organism. Developmental defects associated with T-cell maturation, usually lead to impaired immune responses and vulnerability to infections, or the development of autoimmune diseases. The microenvironment of the thymus not only contributes to the development of those cells, but also eliminates T-cells that could recognize and damage self-cells even in the absence of a virus infection or another problem (e.g. malignant transformation of the cell) (Gordon et al., 2011).

Thymopoiesis needs a constant source of Hematopoietic stem cells (HSCs). An already primed T-cell progenitor can enter the thymus via the bloodstream. After its entrance to the thymus it undergoes a series of adequately studied maturation steps as depicted below :

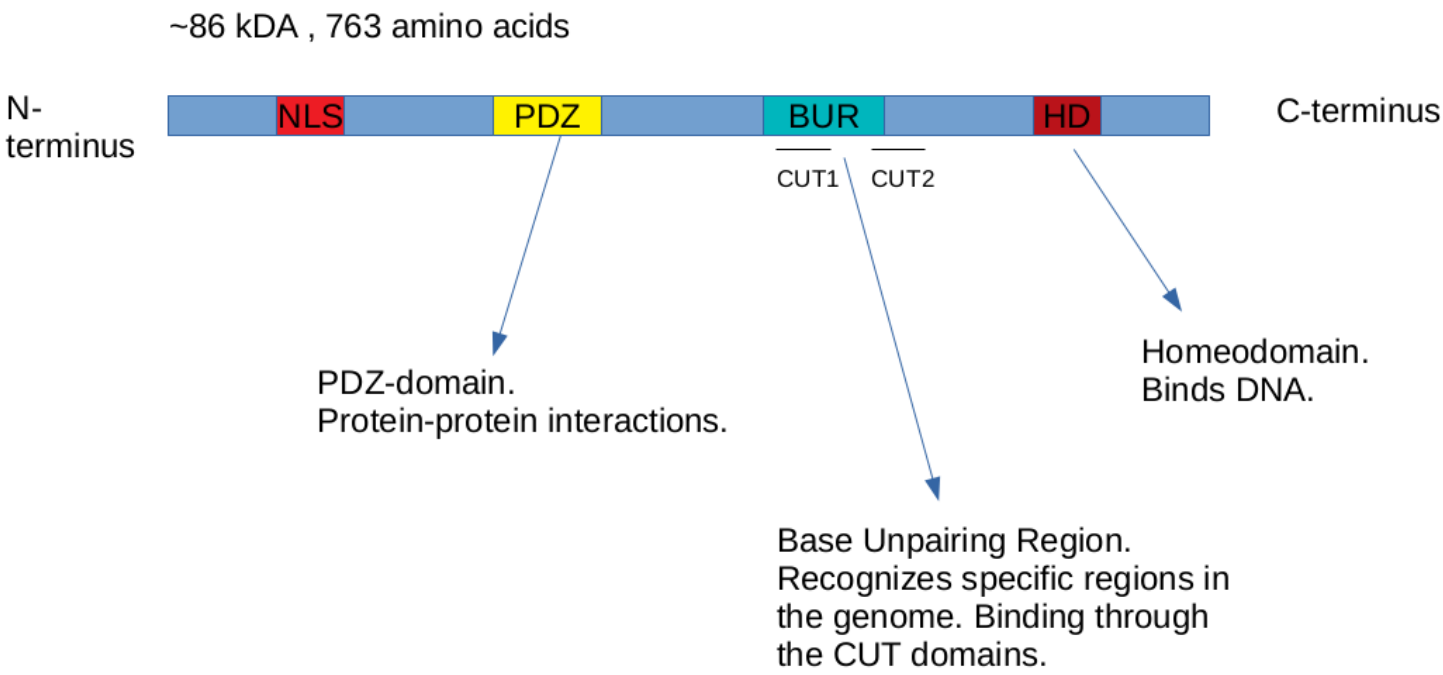


Developing a new paradigm for thymus organogenesis. C. Clare Blackburn et al. Nature Immunology Reviews, 2004.

A crucial developmental step is the Double Positive stage (DP from now on). This stage is characterized by the presence of two surface glycoproteins, CD4 and CD8. Prior to this stage, the cells have undergone a selection process : Cells that can't produce a functional TCRβ-chain undergo apoptosis, while the remainder start producing simultaneously the two gylcoproteins. The T-cell receptor, a surface protein found in T-cells that mediates antigen recognition (thus mediates immune specific responses directed towards a ''foreign'' molecule), is made up from two extremely variable α and β chains (in 95% of the cases) and other accessory invariant protein molecules. This extreme variation is the result of a process termed VDJ-recombination, a site specific recombination process mediated by specific proteins (RAG1 and RAG2) and other ubiquitously expressed DNA-repair proteins. While TCR-β recombination is a requirement for the entrance to the DP stage, TCRα recombination initiates after the activation of the CD4 and CD8 genes.

During the DP stage, the cells are tested for a functional TCR that can bind major histocompatibility complex ligands (present in the surface of other cells in the thymus, like epithelial cells) with intermediate affinity. If a cell cannot bind with enough affinity an MHC-ligand, then it cannot carry out effectively its desired action and it is of no use. On the other hand, a cell that recognizes an MHC-molecule with extreme affinity could damage healthy cells. Cells that manage to mature successfully will have to commit to a specific lineage : Either CD4+CD8- cells or CD4-CD8+ cells, which carry out distinct biological functions. After this commitment the cells leave the thymus (Koch et al., 2011) (Krangel, 2009).

Starting from the DP stage, the expression of the *Satb1* gene can be detected. The *Satb1* gene, codes for a nuclear protein needed for the correct maturation of T-cells (Alvarez et al., 2000) . The structure of the SATB1 protein is shown below :



Source: Atlas of Genetics and Cytogenetics in Oncology and Haematology

SATB1 was originally isolated for its potential to recognize and bind to a Matrix-attachment DNA region (MAR from now on) in the *Igh* gene enhancer. MARs are thought to tether chromatin to the nuclear matrix, a polymeric structure found inside the nuclei (Dickinson et al., 1992). The team that isolated SATB1 also noticed that it could bind AT-rich regions : SATB1 could bind ATC – rich clusters that were depleted of guanines. They finally proposed that SATB1 couldn't come in contact with the actual bases : Rather it seemed that it could recognize the DNA-shape of those sequences (Dickinson et al., 1992). SATB1 possesses specific domains for DNA binding shown in the above picture (CUT1, CUT2, Homeodomain) (Sunkara et al., 2018).

SATB1 also contains a PDZ domain which allows homo polymerization of the protein and the establishment of interactions with other proteins. Although a typical activation or repression domain is not evident, it has been shown multiple times that SATB1 may alter the expression of several genes in different contexts (Sunkara et al., 2018) (Cai et al., 2006). A logical explanation for this capability is the interaction of SATB1 with chromatin modulators. Indeed it has been shown that SATB1 interacts with HDAC1 and functions as a repressor when it is phosphorylated at serine 186 by the protein PKC (Kumar et al., 2006).

Early work has shown the importance of SATB1 for the homeostasis of T-cells in the thymus. A global *Satb1* KO mouse revealed that the loss of *Satb1* leads to an extremely reduced lifespan. Moreover the same study pointed out various abnormalities in the thymus : The thymus was smaller and the number of thymocytes was reduced. The T-cells seemed to be arrested at the DP stage and this arrest was attributed to the deregulation of numerous developmental-associated genes. In a late study, evidence that SATB1 directly regulates the expression of the *Rag1* and *Rag2* genes was presented, explaining partially the observed block (Hao et al., 2015). Finally, the few CD4 single positive T-cells that managed to leave the thymus and enter the periphery seemed to be unable to respond properly to stimuli (Alvarez et al., 2000).

A landmark paper in 2006, published by the team of Shigematsu, proved that SATB1 may mediate regulation of target genes via the creation of chromatin loops (Cai et al., 2006). It was shown that the coordinated expression of various cytokine genes, needed for the proper activation of the Th2 T-cell subset, needs the formation of chromatin loops mediated by SATB1. The gene regulation model is depicted below :

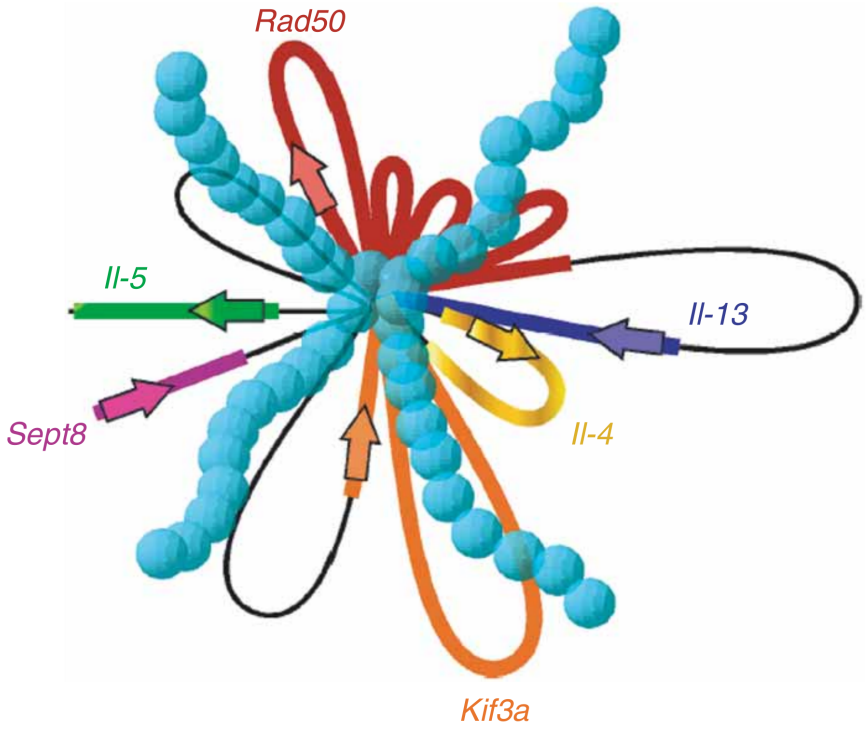
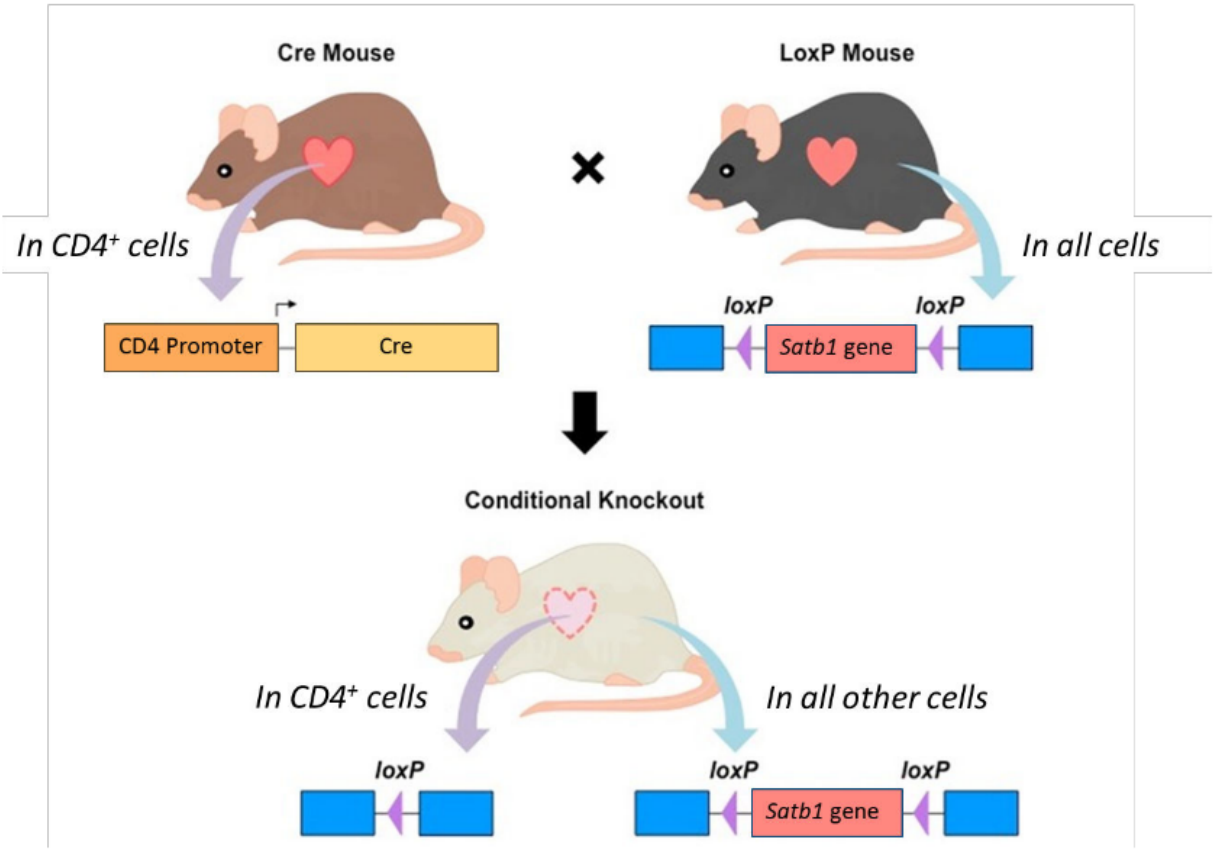


Image source : Cai S. et al., 2006
SATB1 is illustrated as blue circles.

In the aforementioned model, an intron of a lowly expressed gene, *Rad50*, operates as an enhancer element for multiple cytokine genes that are located on the same chromosome. SATB1 forms ‘’bridges’’ between the promoters of the cytokine genes and the enhancer element. Worth noting is the fact that this interaction can occur thanks to the binding domains of SATB1 and its capability of homo - polymerization. In the absence of SATB1, these loops are not formed and the proper activation of Th2 cells is impaired (Cai et al., 2006).

SATB1 is expressed in a highly specific tissue manner. Besides T-cells, its expression can be observed among others, in the brain (Balamotis et al., 2012) and in embryonic stem cells (Agrelo et al., 2009) carrying out important developmental programs. The global *Satb1* KO mice studied at 2000, exhibited various neurological defects, pointing out that the premature death of these mice may be a cause of abnormalities in the central nervous system (Alvarez et al., 2000). Finally, abnormal expression of SATB1 in malignant cells has been correlated with aggressive cancers and poor prognosis (Pan et al., 2016), indicating the tight need of regulation for the *Satb1* gene.

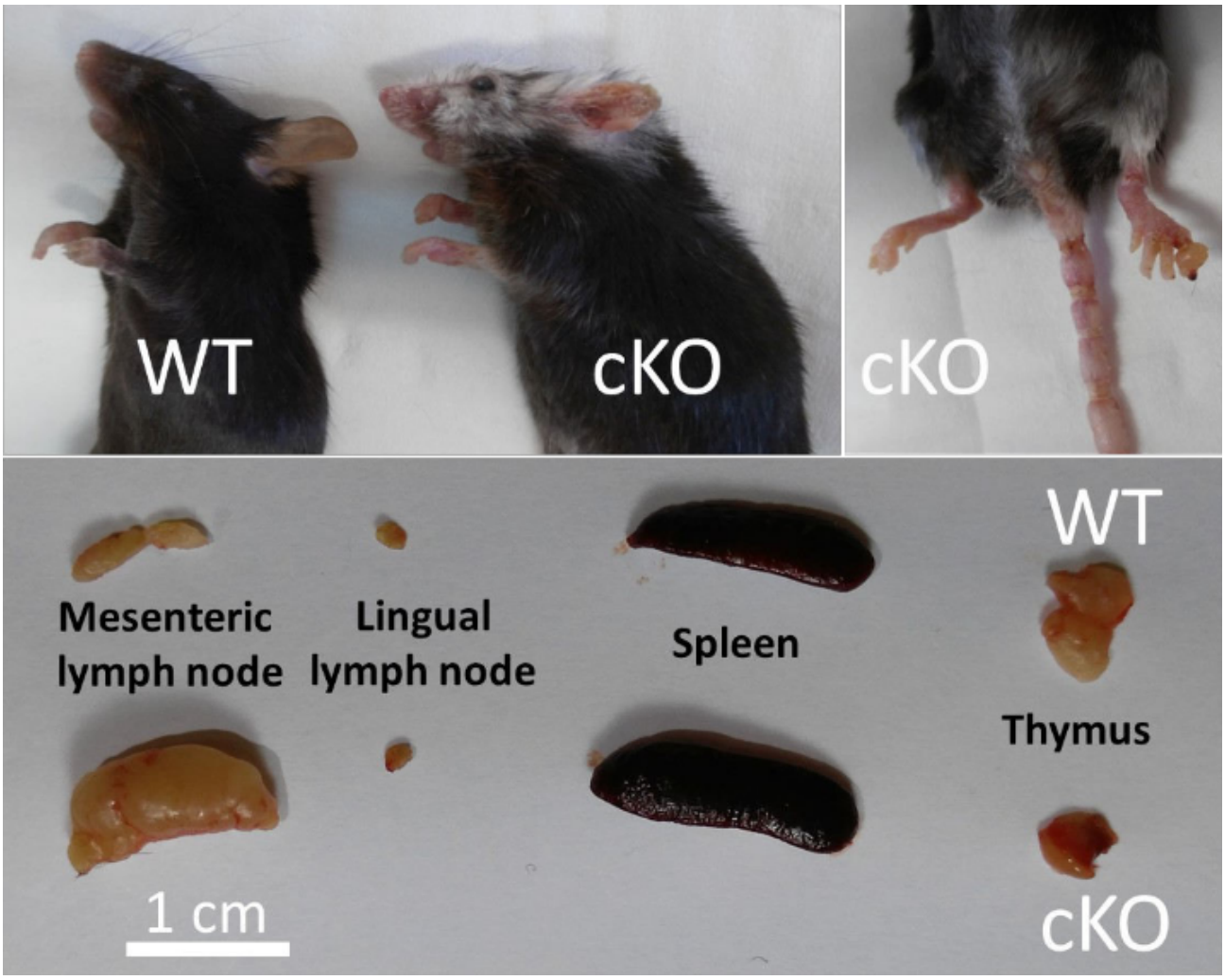
The lack of a thymocyte specific conditional *Satb1* knockout mouse has limited the possibility to study its effects in the immune system in vivo. Since global *Satb1* knockout animals exhibit neurological defects and die very early, it is not possible to determine safely the consequences of its ablation in thymocytes. For that reason a conditional *Satb1* knockout mouse strain was bred in the Spilianakis lab as shown below (unpublished data).



Genaration of a conditional *Satb1* knockout mouse line. Image created by Tomas Zelenka.

A mouse expressing the Cre recombinase under the *Cd4* promoter (thus expressing the Cre recombinase in T-cells starting from the DP stage) was crossed with a mouse having two loxP sites between the third exon of the *Satb1* gene. The Cre recombinase's action will ultimately remove sequences found between two loxP sites, if they are correctly oriented upon the sequence (Zheng et al., 2000). The offspring of the above mouse strains will be unable to produce SATB1 in thymocytes.

The loss of *Satb1* in thymocyes leads to the development of autoimmunity. The conditional knockout animals have a reduced lifespan (unpublished data from the Spilianakis lab). Although they live more than their global *Satb1* knockout counterparts, additional symptoms arise like joint inflammation and an enlarged spleen. The thymus appears to shrink in size as in the original global *Satb1* knockout (Alvarez et al., 2000).



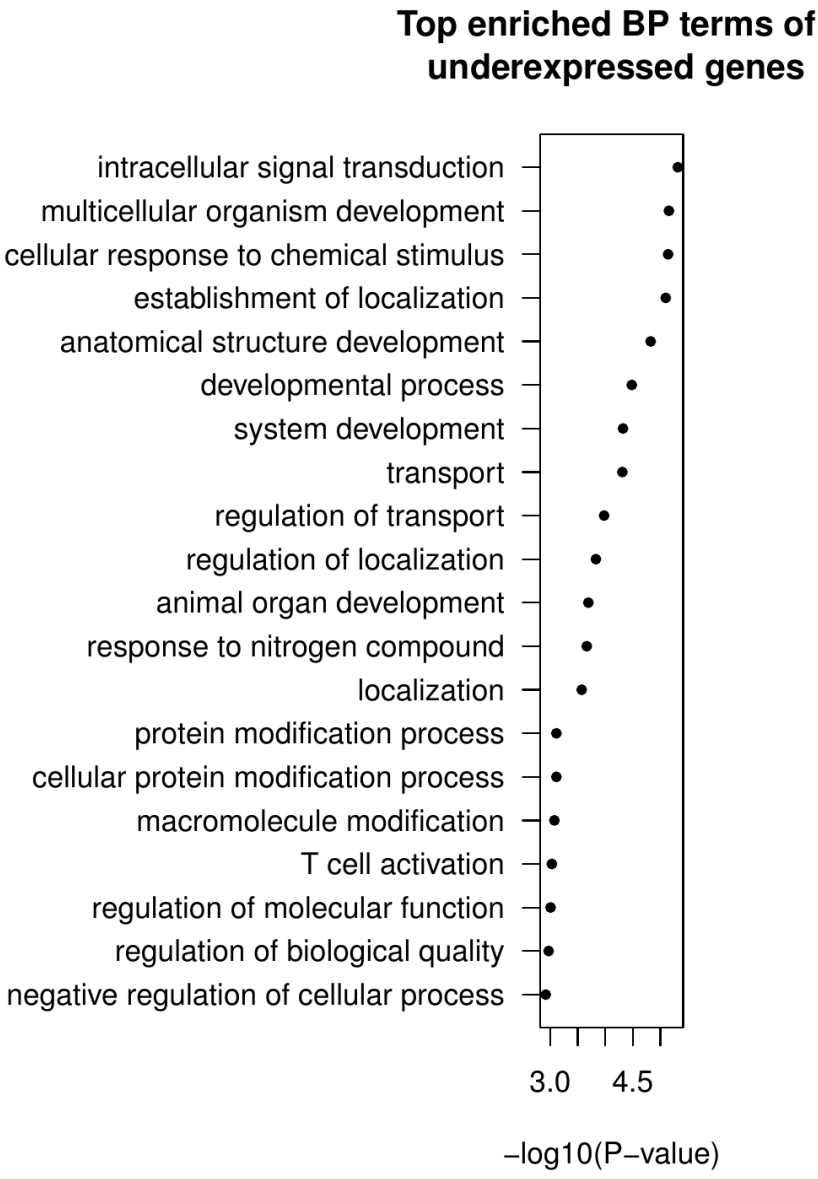
Comparison of various organs of Wt mice and *Satb1* conditional knockout mice. Image created by Tomas Zelenka.

In this study a number of next generation sequencing techniques were employed in order to investigate SATB1's contribution to autoimmunity. Whole thymus samples and sorted DP cells, derived from Wt mice and *Satb1* conditional knockout mice, were used as input material for various next generation sequencing techniques in order to investigate the molecular events regulated by SATB1 in these cells. We show that multiple axes are affected by the absence of SATB1, leading to thymocyte intrinsic defects along with an altered thymus environment. These results point out at various developmental defects and contribute to our understanding of the phenotype.

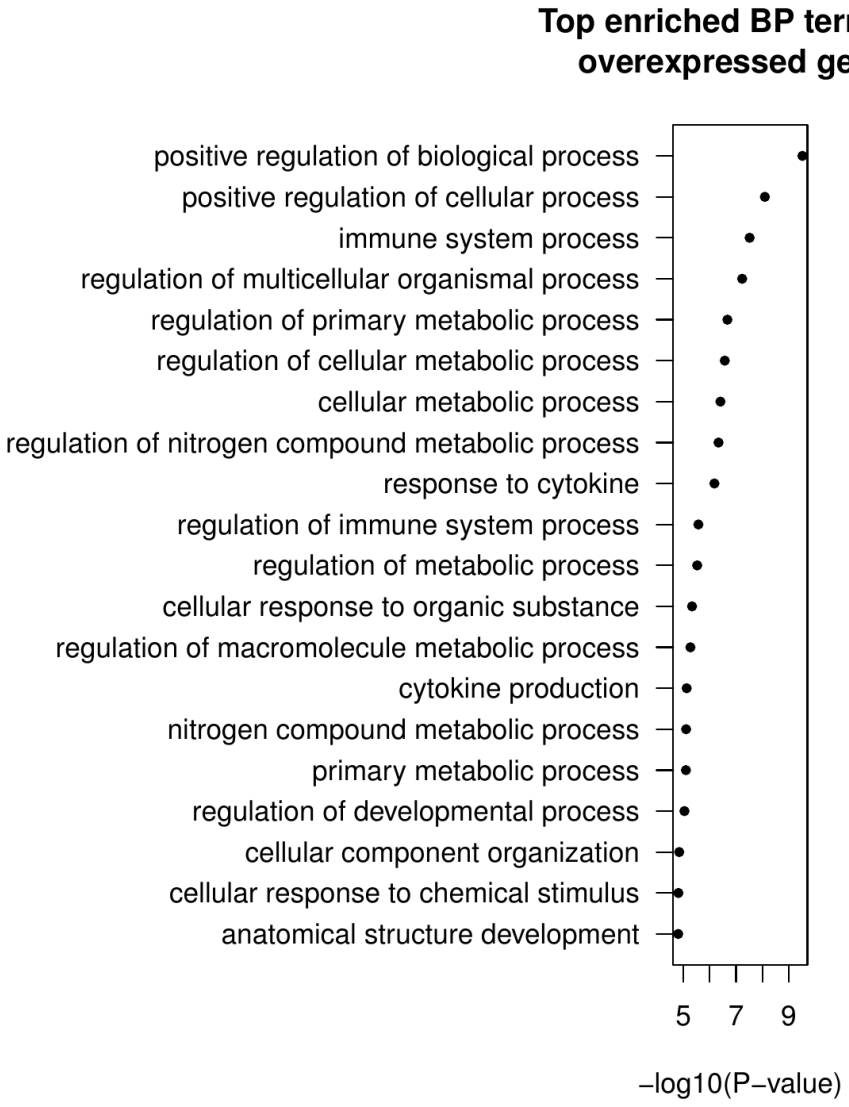
Results

The absence of SATB1 affects the expression levels of multiple genes in thymocytes

Sorted DP thymocytes were used to compare the expression levels of genes between Wt and *Satb1* knockout mice. The RNA-seq dataset was analyzed as described in the "Materials and Methods" section and for the extraction of differentially expressed genes an FC cutoff of 1.5 was used along with a p-value of less than 0.05. 445 genes were deemed as overexpressed, while 374 genes were deemed as underexpressed. Functional enrichment analysis revealed that overexpressed genes were mainly associated with immune-system processes, while underexpressed genes were also associated with more general categories, like signal transduction.



The expression of signal transduction genes seemed to be positively affected by SATB1's presence. Some T-cell related pathways are also present like the "T cell activation" BP term.

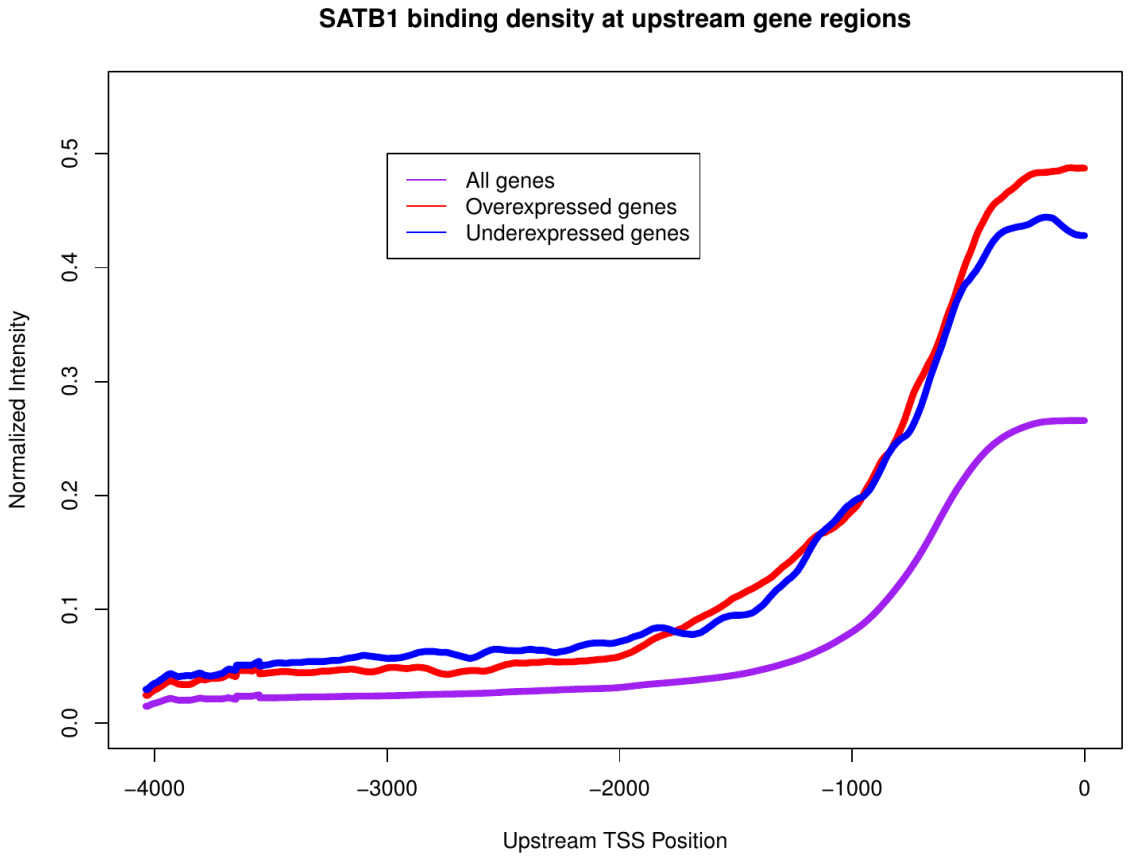


Immune related pathways seem to dominate the functional enrichments observed for overexpressed genes. It seems that genes related to the differentiation of various T cell subsets are enriched. The differentiation of this subset is an event that should happen at the periphery, in later developmental steps. This finding is in accordance with the phenotype of the global *Satb1* knockout mouse (Alvarez et al., 2000), where such events were reported.

SATB1 binds directly the promoters of a large portion of differentially expressed genes

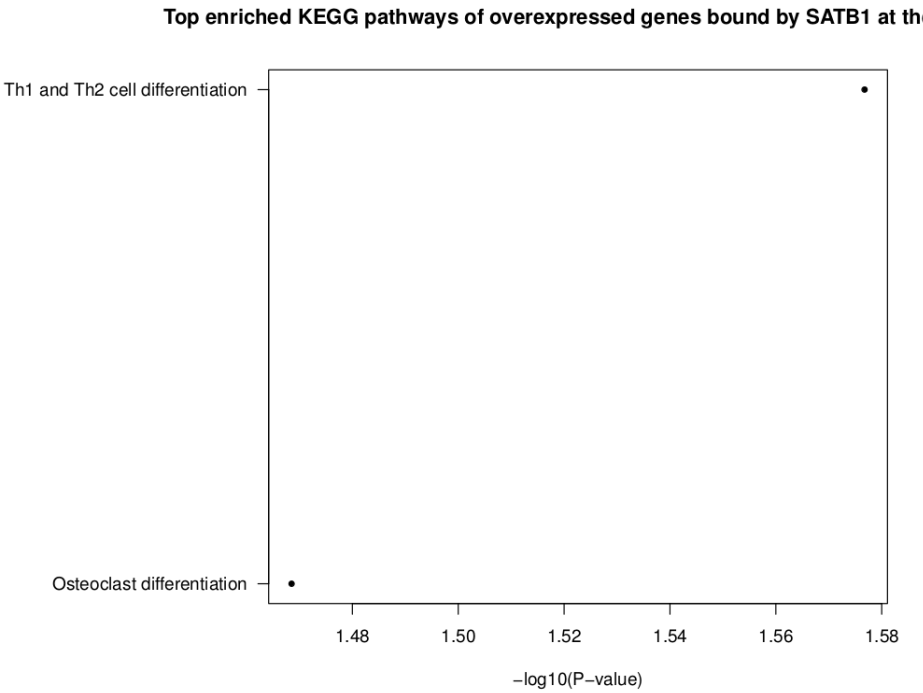
We next investigated whether SATB1 could bind directly the above differentially expressed genes. The binding of SATB1 at the genomic loci of differentially expressed genes could pinpoint genes that are regulated directly by SATB1. We took advantage of a previously published Chip-seq dataset for SATB1 performed on whole thymus extracts (Hao et al., 2015).

Average gene profile plots were created for upstream and downstream regions of the genomic loci of differentially expressed genes. The plots revealed that SATB1 directly binds a large amount of differentially expressed genes at their promoter regions, suggesting possible direct regulatory events .

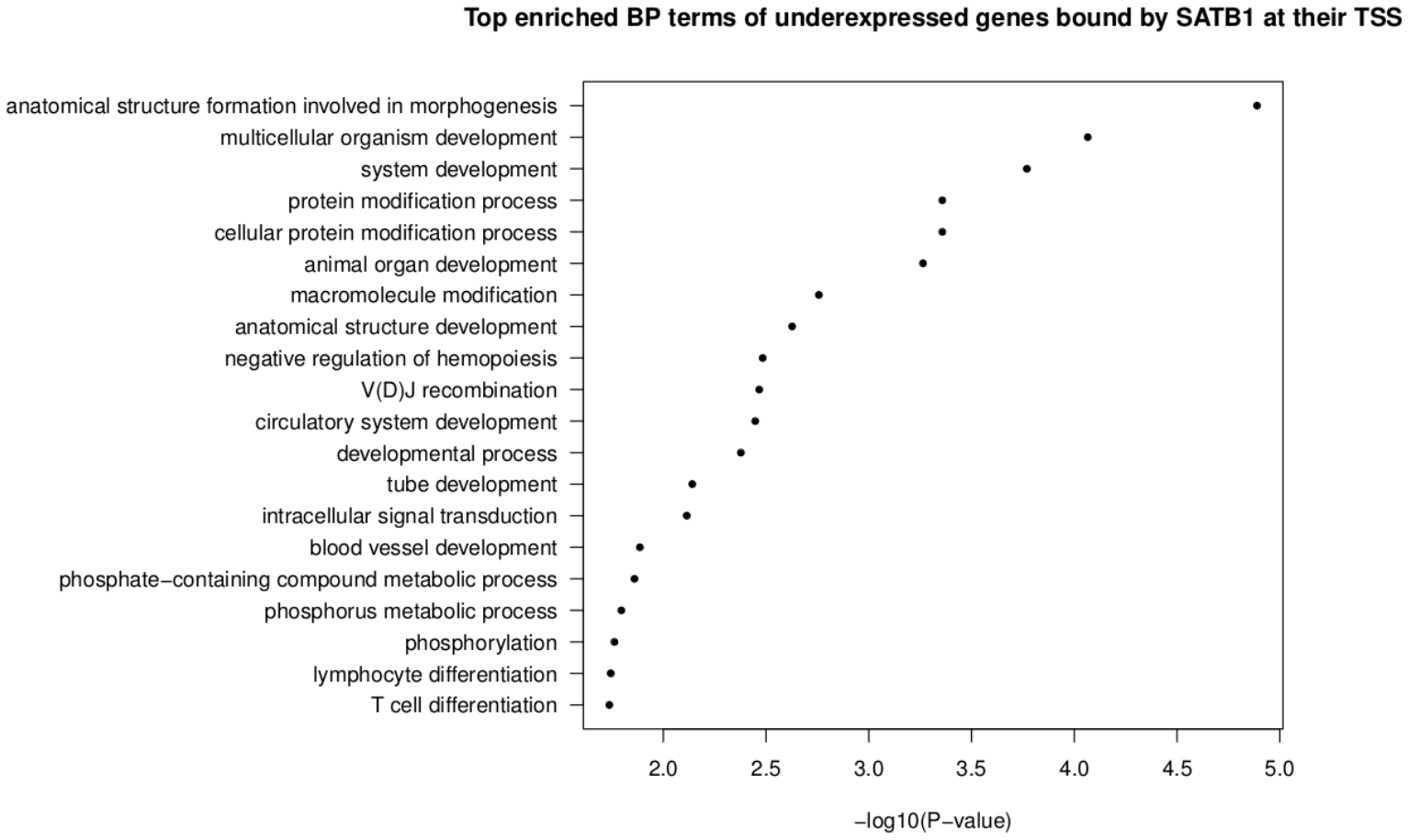


It is also evident that SATB1 binds the promoters of a large number of genes. Around twenty percent of the genes, irrespectively of them being expressed or differentially expressed, seem to be bound by SATB1 in proximity to their promoter sites. These results are in accordance with the known events SATB1 orchestrates in thymocytes.

Differentially expressed genes that were bound upsteam of their transcription start site by SATB1 (800 bp upstream - TSS) were isolated. 230 overexpressed genes were bound by SATB1 at their transcription start site, while 179 underexpressed genes exhibited this behavior. Functional analysis was performed for these subsets.



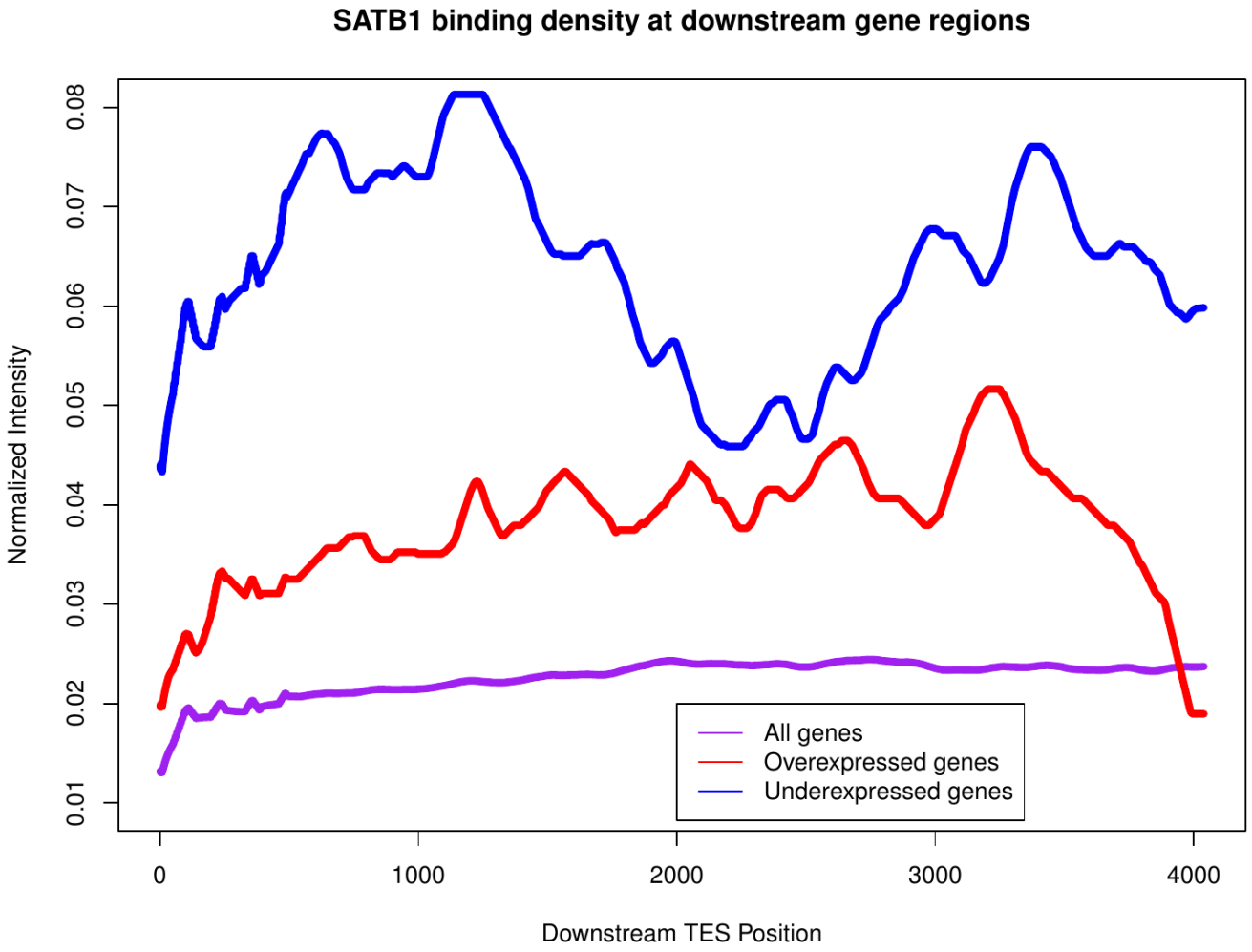
The genes falling in the KEGG pathway “Th1 and Th2 cell differentiation” are *Gata3, Ifngr1, Stat1, Il2ra, Jun* and *Maf. Gata3* is a crucial transcription factor needed for Th2 lineage commitment of CD4+ T-cells in the periphery, that also plays a role in post-selection double positive thymocytes (I-Cheng Ho et al., 2009). In general it seems that genes that play a crucial role for the final differentiation of peripheral CD4+ cells to the Th2 subset, are de- repressed in the absence of SATB1 while they were originally bound by it. This could point out to a model where SATB1 represses directly genes that are important players in later stages of T-cell differentiation.



Developmental related genes have lower expression values in the *Satb1* knockout thymocytes and are simultaneously bound by SATB1 at their promoter sites. A signal transduction term once again appears and includes the *Pten* gene, a crucial phosphatase of the *Akt* pathway (Zhenbang Chen et al., 2005). The “T cell differentiation” term contains the genes *Tcf7, Bcl6, Satb1, Il4ra, Rag2, Ccr7, Socs1, Egr1, Rag1, Mafb.* The connection between SATB1 and the genes *Rag1* and *Rag2* is already known and it is crucial for the development of DP thymocytes (Hao et al.,2015). Further investigation is needed for the other immune related genes in this term.

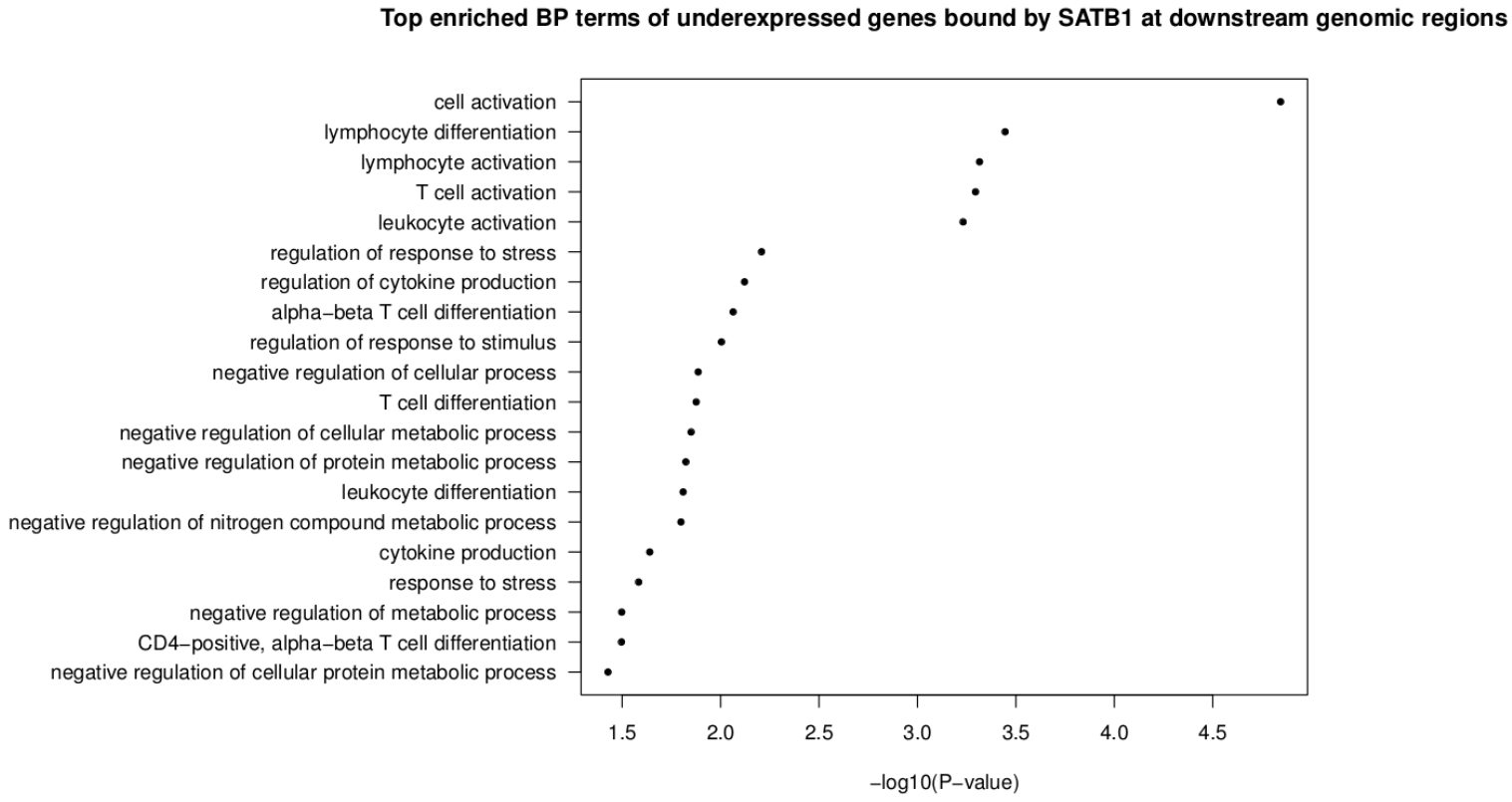
SATB1 preferentially binds underexpressed genes at downstream genomic regions

An average gene profile plot for differentially expressed genes revealed something unexpected.



It is evident that downstream binding events are not common. Just 1% of the total genes exhibited a SATB1 binding event directly after their transcription termination site. However it seems that underexpressed genes are enriched for downstream SATB1 binding, having percentages up to 8% for specific areas (e.g. 1kb downstream of the transcription termination site). This phenomenon was further validated by constructing the same plot only for genes that didn't have a neighboring promoter nearby (less than 10kbp) and the same trend persisted (data not shown).

The underexpressed genes bound by SATB1 at downstream regions were isolated. 90 underexpressed genes were isolated this way. Functional analysis was performed.

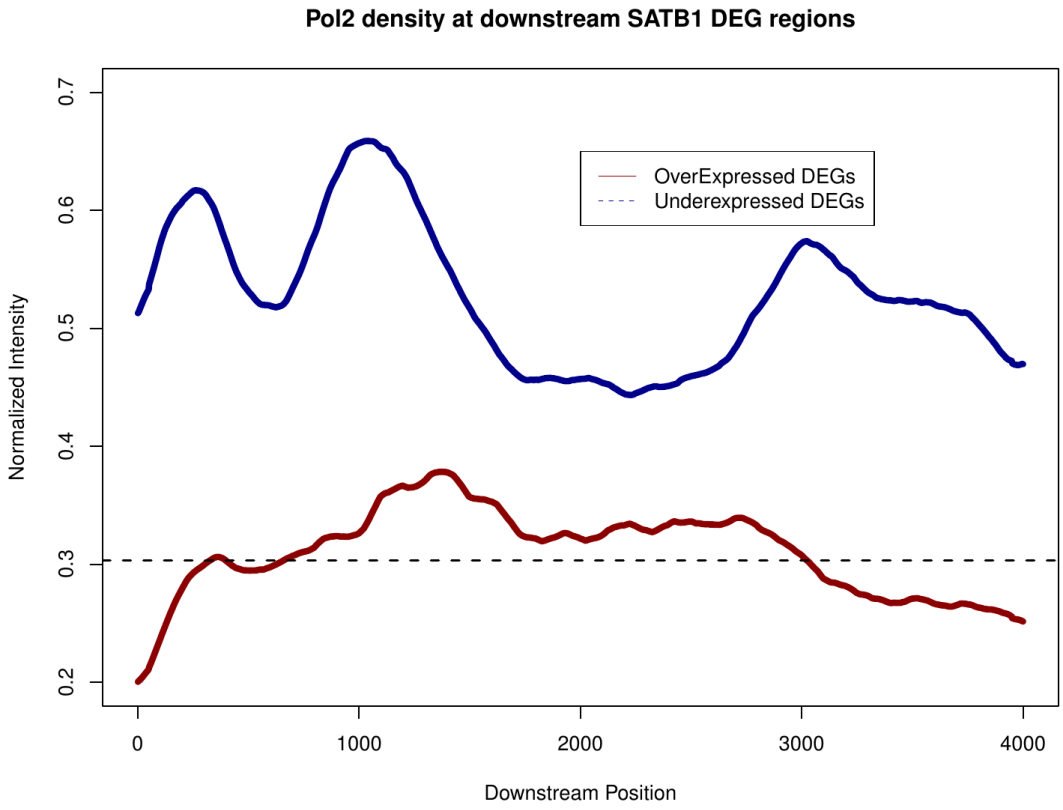


Immune-related biological processes are enriched. The functional significance, if any, of such binding events remain to be elucidated. Genes that fall in the “Cell activation” term are the following : *Tcf7, Rps6ka1, Slamf6, Id2, Rgcc, Satb1, Cd6, Tpd52, Cd2, Lag3, Il4ra, Rag2, Ccnd3, Ly6d, Socs1,Zc3h12a,Rag1,Cnr2,Capn3.*

Worth noting is the fact that SATB1 binding in the downstream genomic regions is usually accompanied by simultaneous binding of SATB1 nearby their transcription initiation site. 53 out of the 90 downstream-bound underexpressed genes were occupied by SATB1 at both of the aforementioned sites. This could suggest an ‘’insulation’’ like effect of SATB1 for these genes : Perhaps binding of both ends of the genomic region of a gene could isolate the gene from its neighborhood. Another potential regulation mechanism involving SATB1 binding at the boundaries of a genomic locus could involve the creation of short chromatin loops in that locus, as discussed below.

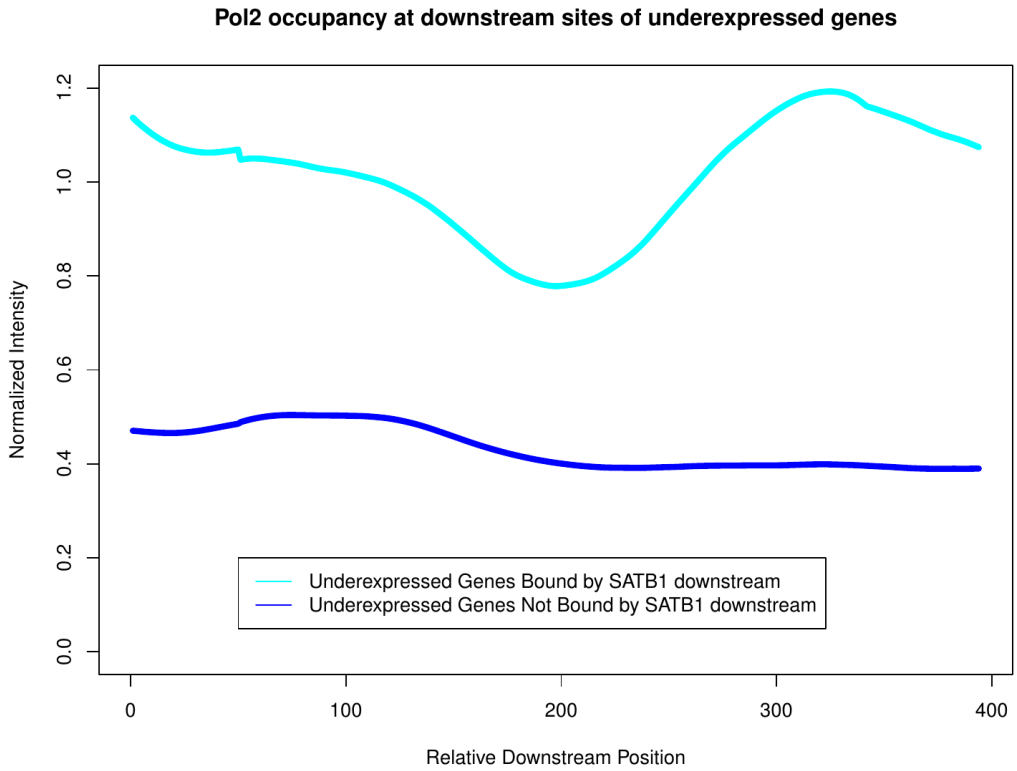
Increased Pol2 occupancy at genomic regions downstream of underexpressed genes

The above finding prompted us to look how POL2 occupancy is distributed across the downstream genomic regions of underexpressed genes utilizing a publicly available Chip-seq dataset (ENCFF001LSR accession number from ENCODE). This occupancy could reveal possible regulatory events at these regions. Average gene profiles were constructed.



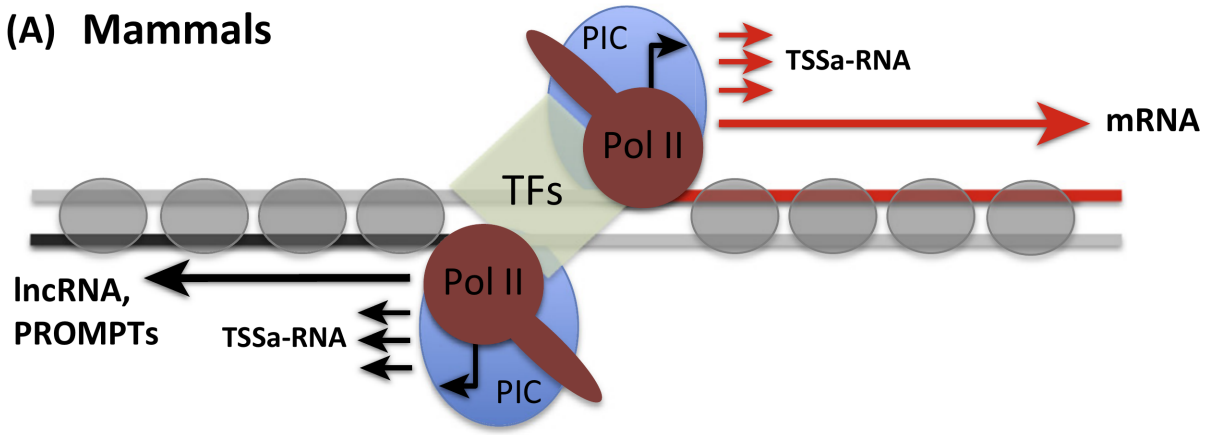
The dashed line indicates the average occupancy signal of POL2 across the same regions for all of the genes. It is evident that no tendency can be observed for the overexpressed genes, while the underexpressed genes exhibit increased POL2 occupancy at these regions. A much needed control is needed though : The increased POL2 occupancy may be irrelevant to SATB1 binding at the same loci, since average gene profiles depict average tendencies across group of genes.

Another average gene plot was constructed. This time the underexpressed genes were split into two groups : The first group contained all the underexpressed genes that were bound by SATB1 downstream and the other the rest of the underexpressed genes. Average gene profiles for POL2 occupancy were once again constructed for these groups.



It seems that SATB1 binding at these areas is well correlated with POL2 occupancy. Worth noting is the fact that the above POL2 occupancy is not due to proximal promoter regions. The same plots were constructed by filtering out genes who had a proximal promoter of another gene near their transcription termination site(<=8kb distance) and the trends persisted (data not shown).

A possible model linking the above findings together is the following. SATB1 could mediate short chromatin loops at a subset of genes, between their transcription start sites and their transcription termination sites. This conformation has been proposed to allow rapid POL2 "recycling" after a round of transcription, leading to an overall increase in the transcription levels of a gene. Thus the ablation of SATB1 could lead to the disruption of the loop, leading to the decrease of the expression levels of the corresponding gene. This model could explain why this tendency is only evident for underexpressed genes. The model is depicted below (Pawel Grzechnik et al., 2014).

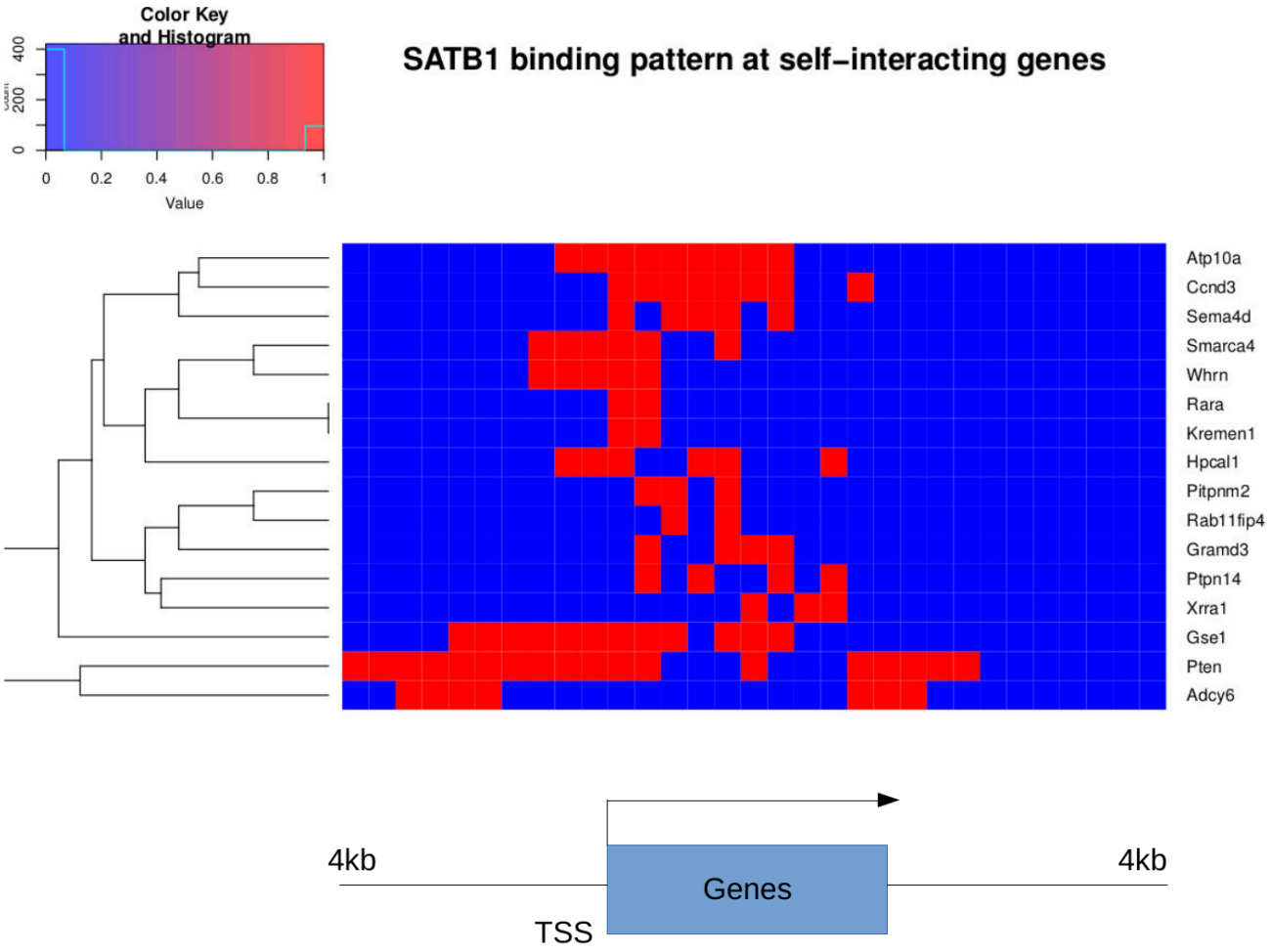


Self interacting genes via SATB1 associated loops are enriched for underexpressed genes

In order to validate the above ''self-looped gene'' hypothesis, the SATB1 HiChIP experiment results were utilized. A set of SATB1 associated loops were retrieved as described in the "Materials and Methods" section. The loop-anchors were 5kb bins.

Genes that showed overlap with a loop anchor were isolated. Furthermore, genes that were found simultaneously in both loop anchors of a loop, were classified as "self interacting" genes. A total of 146 genes were isolated. Of those, 16 genes were simultaneously underexpressed and ''self-interacting''. A permutation analysis revealed that this overlap is significantly enriched : The expected overlap from the permutation analysis showed that on average we would expect an overlap of one gene only.

The 16 genes were then tested for SATB1 occupancy across their genomic regions. A heatmap was constructed and is shown below.



A binding event is illustrated as a red square. Although all of the isolated genes were bound by SATB1, the binding pattern didn't come in accordance to the proposed model : Only *Pten*, *Adcy6* and *Ccnd3* exhibited simultaneous SATB1 binding at the boundaries of their genomic locus. Moreover, only *Pten* and *Adcy6* are part of the underexpressed group found to be bound by SATB1 downstream. Thus, the depicted genes cannot adequately explain adequately the SATB1 binding occupancy at downstream genomic regions of underexpressed genes. Nevertheless further study through biochemical means for the *Pten* and *Adcy6* could reveal rare, as it seems, cases of ''self looped'' genes regulated by SATB1.

SATB1 binding hot-spots are almost always found proximal to gene loci

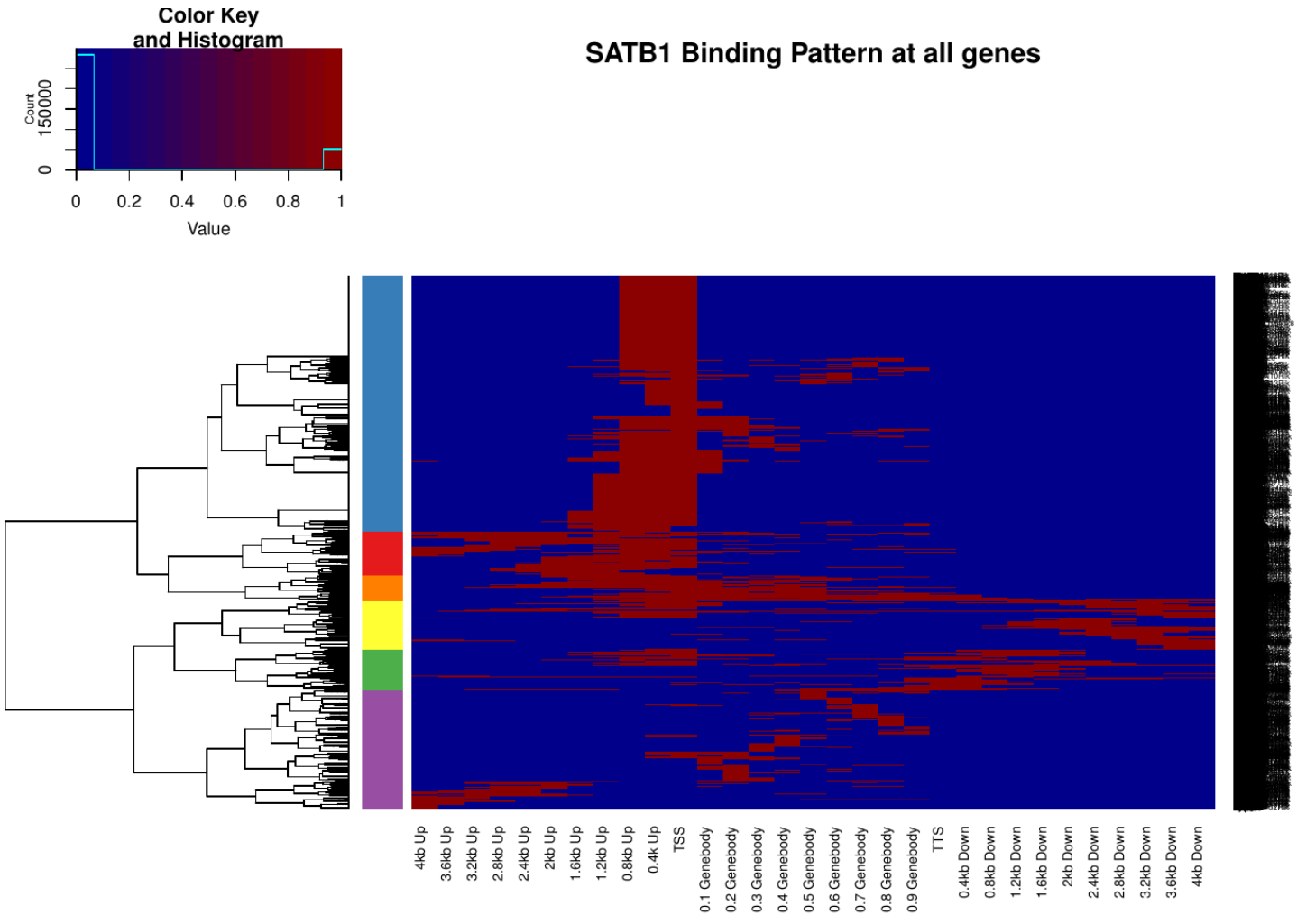
After studying the interplay between the binding events of SATB1 across differentially expressed genes, a different strategy was employed in order to detect, in an unbiased fashion, areas in the genome that showed excessive SATB1 binding. The methodology employed is described in the "Materials and Methods" section and managed to identify 561 excessively bound by SATB1 genomic regions that were larger than 100kb each. Out of the 561 regions only 32 didn't overlap with any gene.

An interesting finding is that the top scoring SATB1 binding hot-spot is found in the genomic locus of *Satb1* itself. This result could suggest a self-regulatory positive feedback loop, explaining in part the very high expression levels of SATB1 during the DP stage of T-cells. Another interesting loci found at the top hot-spots identified, is the *Cd8a* locus. This result could point out at a possible regulatory role

for *Satb1* for an extremely crucial gene during T-cell development. Indeed the expression levels of *Cd8a* exhibit a statistically significant mild decrease (~20% reduced expression) in the *Satb1* knockout DP thymocytes.

Distinct patterns of SATB1 binding across all gene loci

Since the previous analysis pointed out that SATB1 is mainly found in the vicinity of the genomic loci of genes, further analysis of these areas was carried out. A heatmap depicting the occupancy of SATB1 across all the genes bound by it was created. In total 12092 genes were bound at least once at their genebody region or at their upstream (-4000 bp – TSS) or downstream (TES- 4000 bp) region.



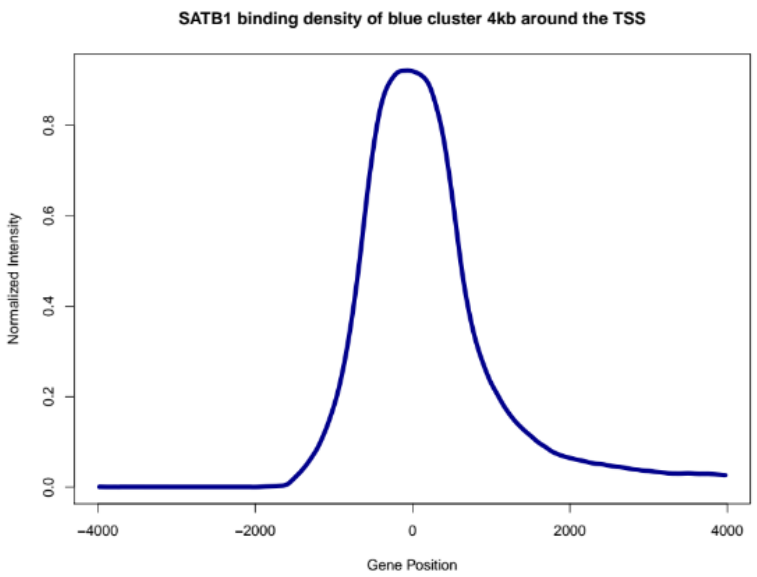
Hierarchical clustering was also performed and the final clusters are depicted on the heatmap with different colors (the number of clusters was not chosen randomly, see “Methods and Materials”). It seems that the majority of genes bound by SATB1, are bound at their promoter region. A well positioned binding region occupying the transcription start site of the genes, is the characteristic of the blue cluster. The red cluster shows a similar behavior, except that binding events also occur at more upstream regions. The orange cluster consists of genes that exhibit SATB1 binding across their whole genebody. The yellow cluster is made of genes showing downstream binding events. Worth noting is that some of the genes in the yellow cluster also show a simultaneous binding at their transcription start

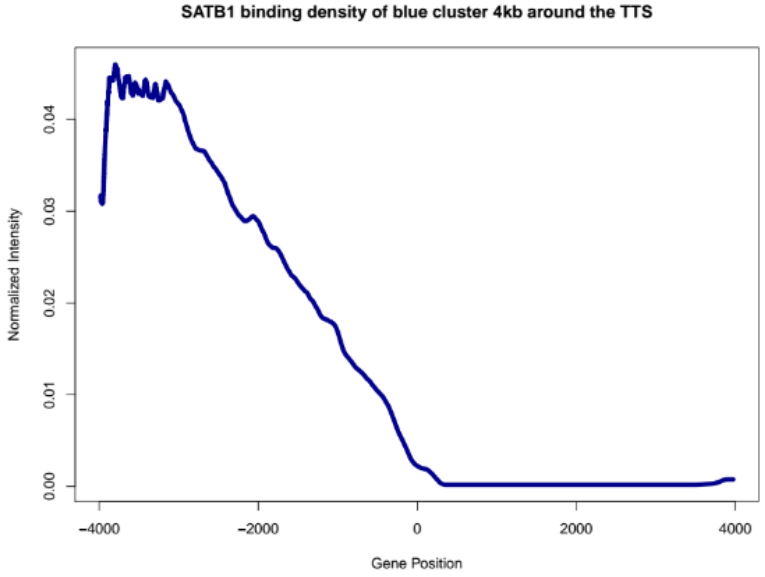
site. The green cluster is similar to the yellow cluster with the exception of downstream binding sites that are more proximal to the transcription termination site. The purple cluster is the more characteristic one though : It contains a large number of genes and it seems that the binding of SATB1 is positioned across the genebody, with no occupation of the transcription start site.

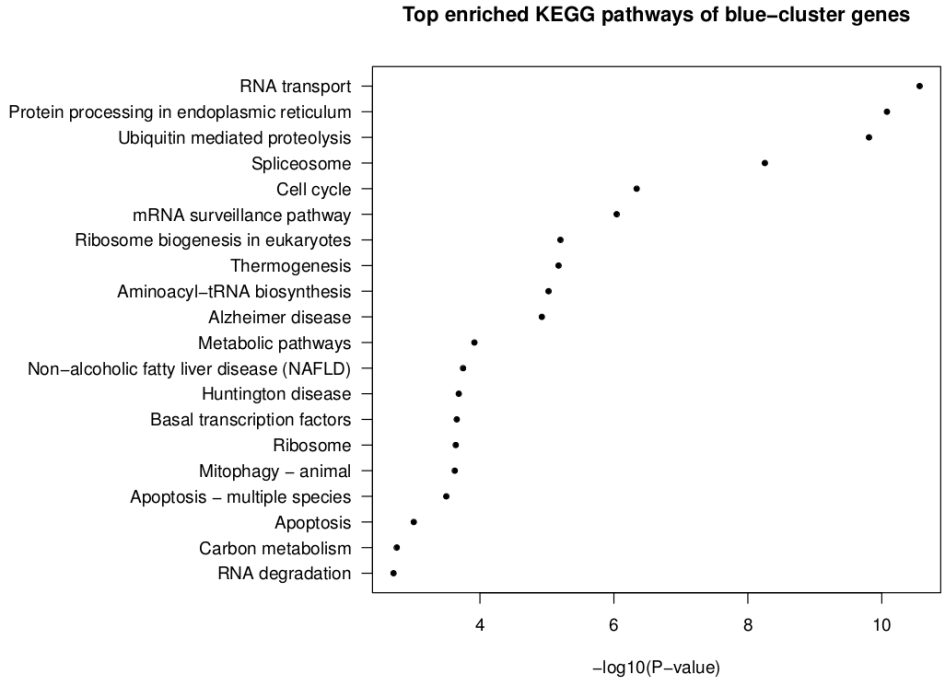
Average gene profiles were constructed separately for each cluster. Moreover the percentage of active genes for each cluster was calculated. Finally functional analysis was performed for each cluster separately.

Cluster #1 Blue

The blue cluster in enriched in necessary - “housekeeping” pathways. Binding is centered around the transcription start site of genes, with minimal binding occurrence in other areas. The genes of this cluster are in the vast majority active.





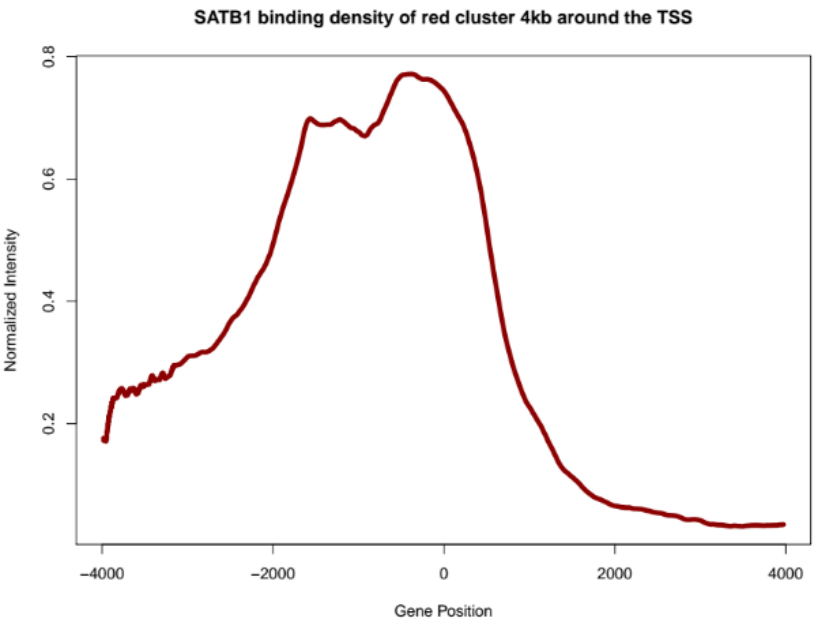


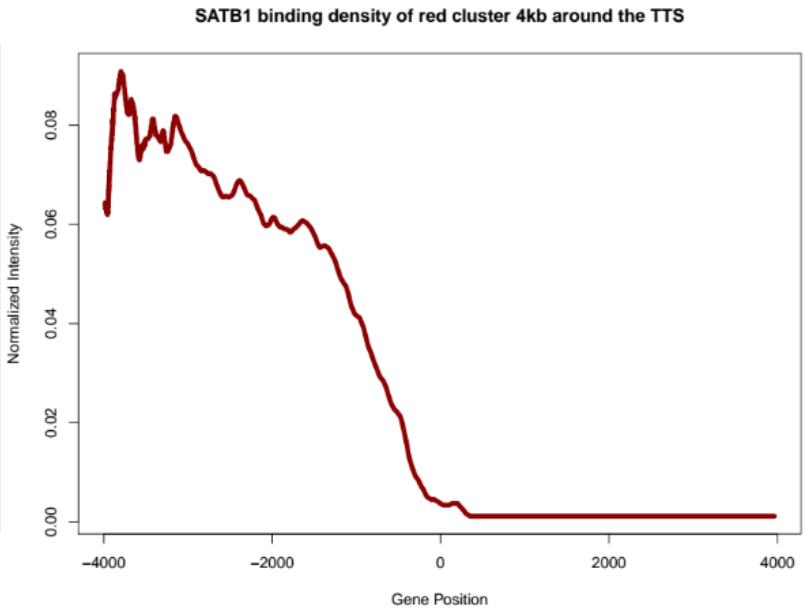
The blue cluster contains **5174** genes

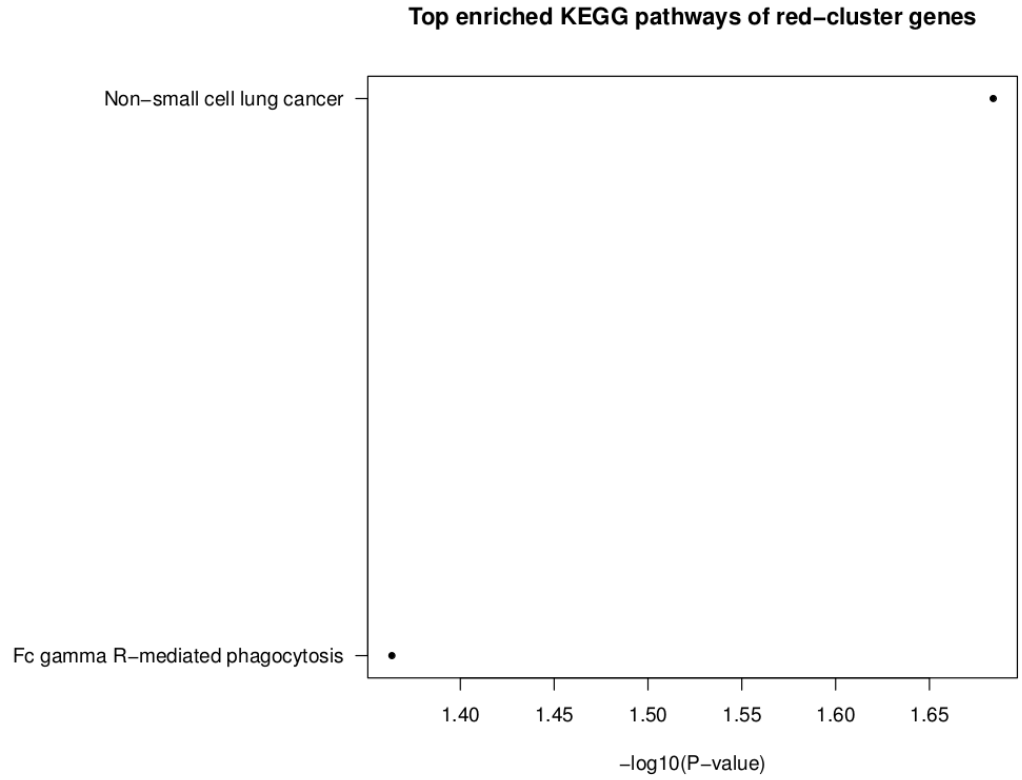
8% of the genes are silent based on the RNA-seq dataset (on average 52,5% of all the genes are silent).

Cluster #2 Red

The red cluster is very similar to the blue cluster. The main difference lies at the more widespread binding upstream of the transcription start sites of genes. The genes of this cluster are also in general active.





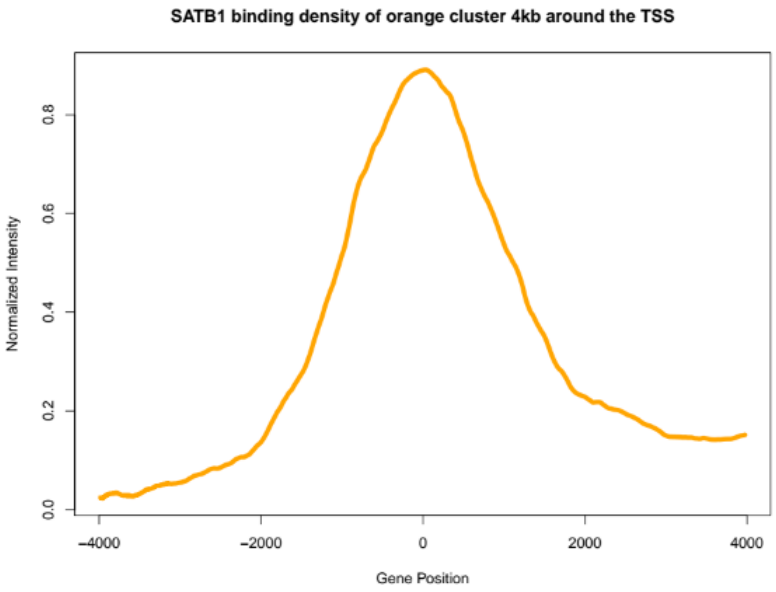


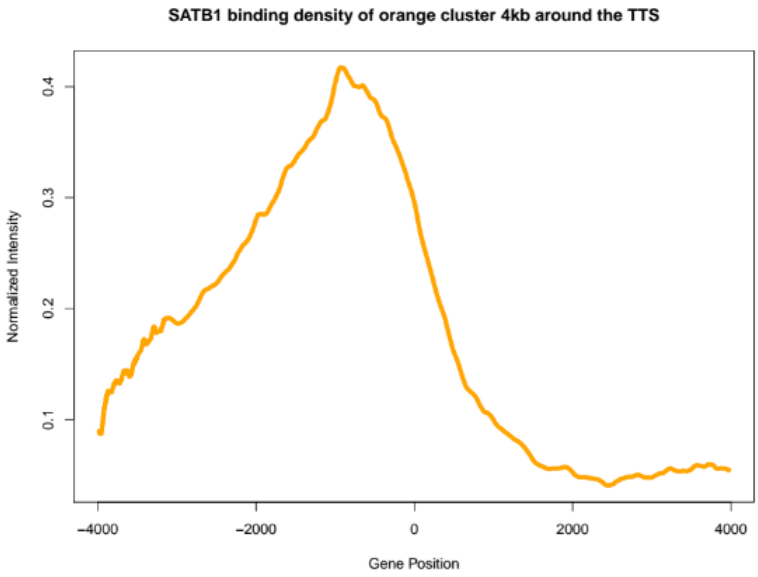
The red cluster contains **893** genes.

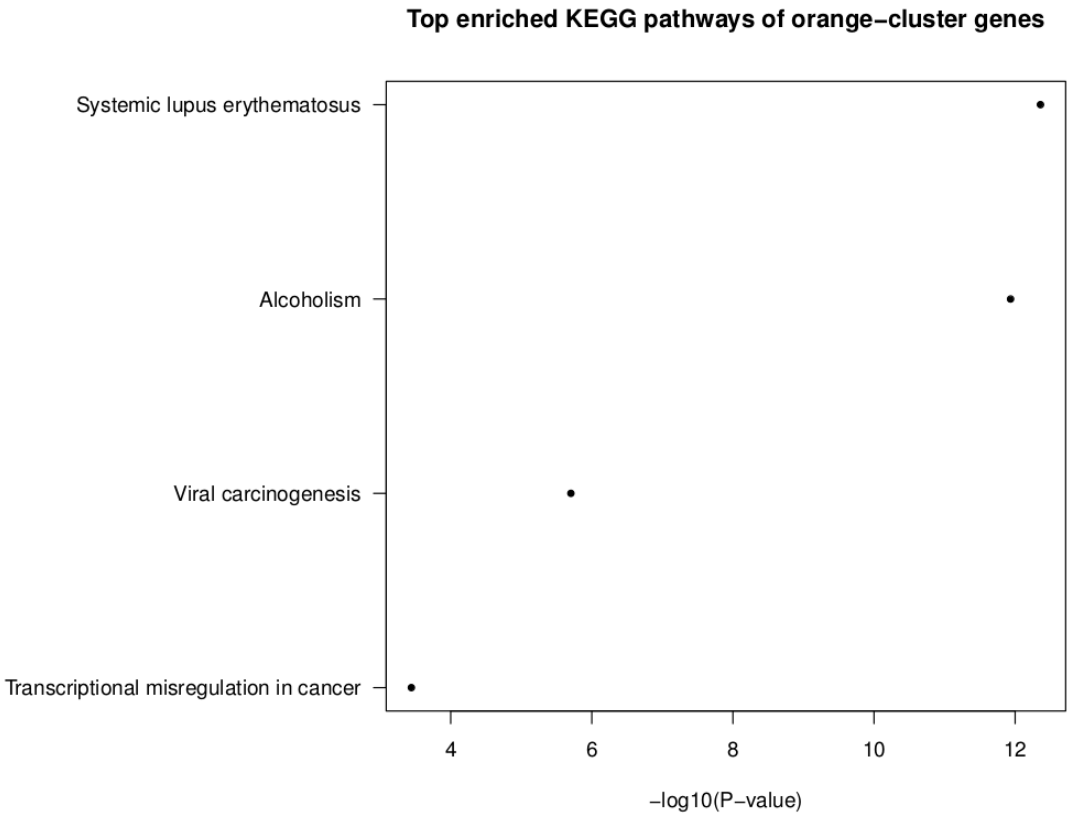
16,7% of the genes are not expressed (on average 52,5% of all the genes are silent)

Cluster #3 Orange

The orange cluster contains a lot of genes that show widespread binding : In a lot of cases it seems that SATB1 binds the whole genebody. A plot showing the overall gene length of the genes found in clusters revealed that the orange cluster contains a lot of genes that are very short (the plot is shown below). These short genes contribute significantly to the overall widespread binding picture. Nevertheless a lot of long genes – members of the cluster (like the *Satb1* locus itself) , exhibit excessive SATB1 binding throughout their genomic loci. The cluster is enriched in immune-related pathways.





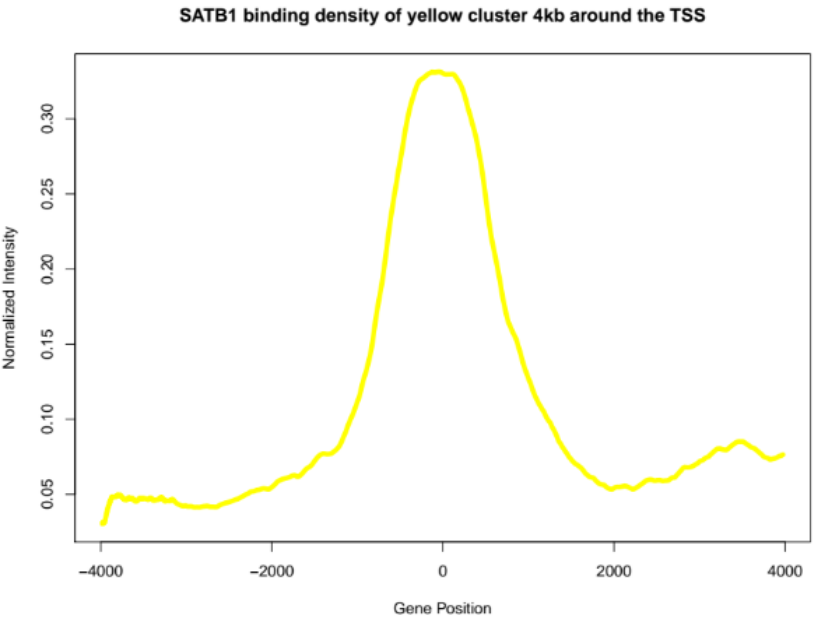


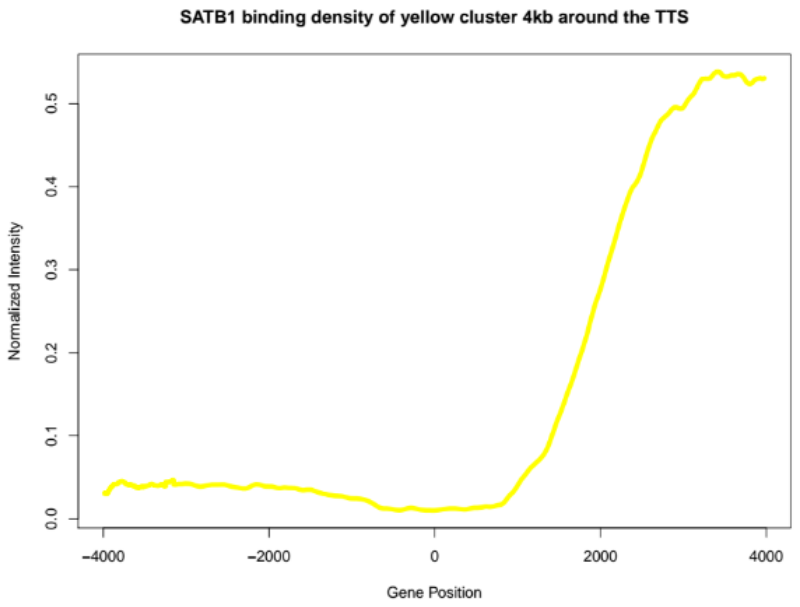
The orange cluster contains **525** genes.

18,09% of the genes are not expressed.

Cluster #4 Yellow

The genes of the yellow cluster show binding tendencies towards their transcription start site. These tendencies are not as high as observed in the previous clusters. Finally these genes are characterized by the presence of a SATB1 binding downstream their transcription termination site.

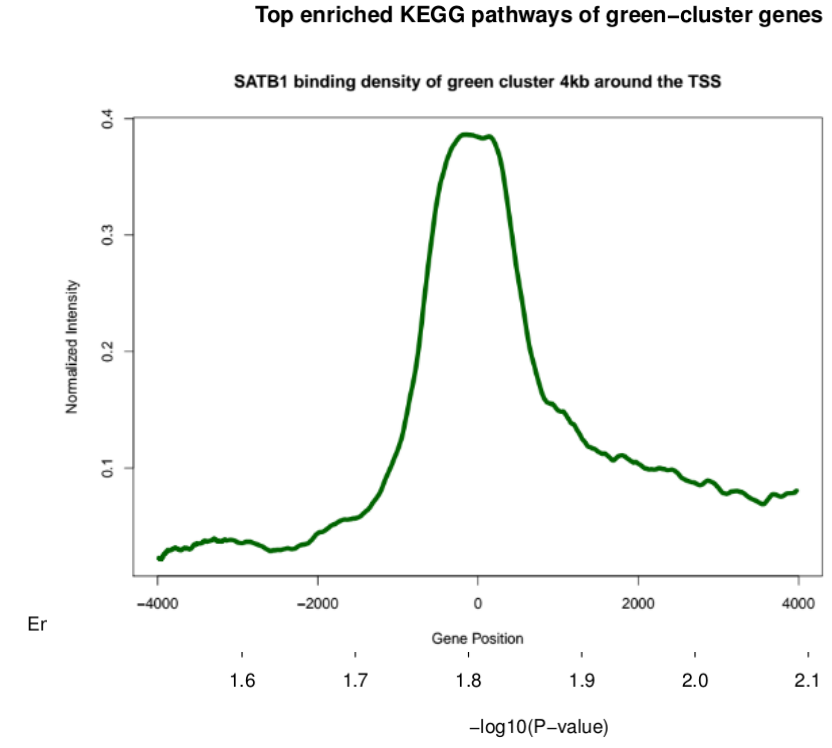


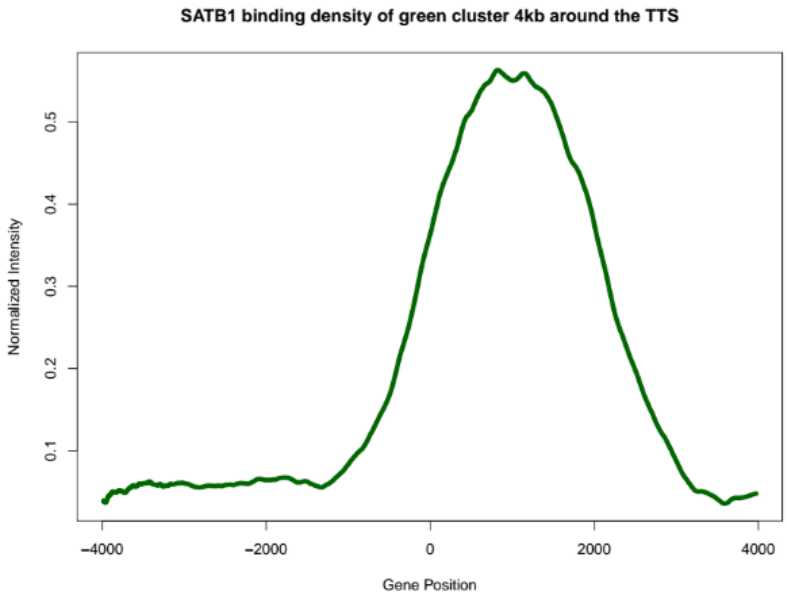


The yellow cluster didn't exhibit any KEGG pathway enrichments. It contains **983** genes. 33% of the genes were not expressed.

Cluster #5 Green

The genes of the green cluster are characterized by the presence of SATB1 near to their transcription termination site.



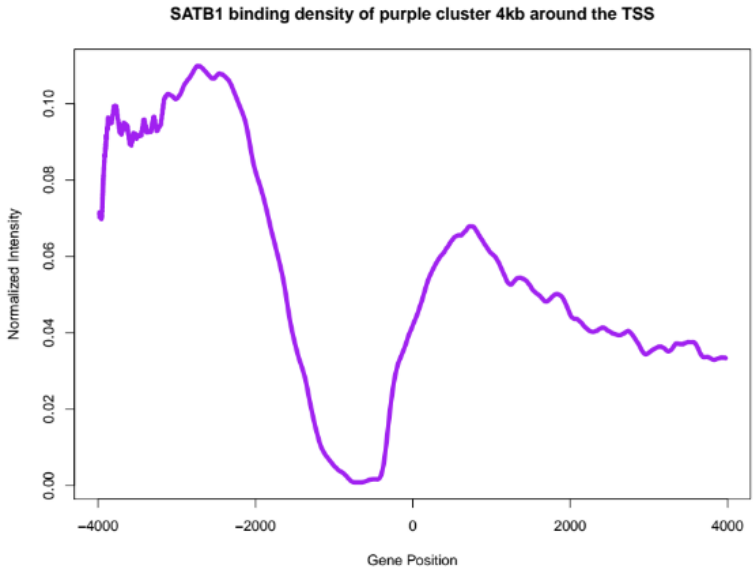


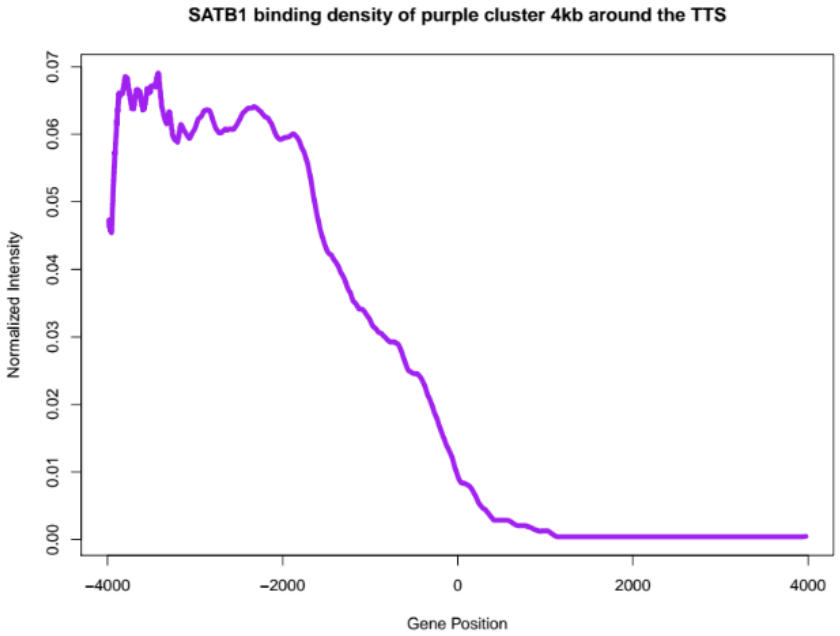
The green cluster contains **807** genes.

32,5% of the genes are not expressed.

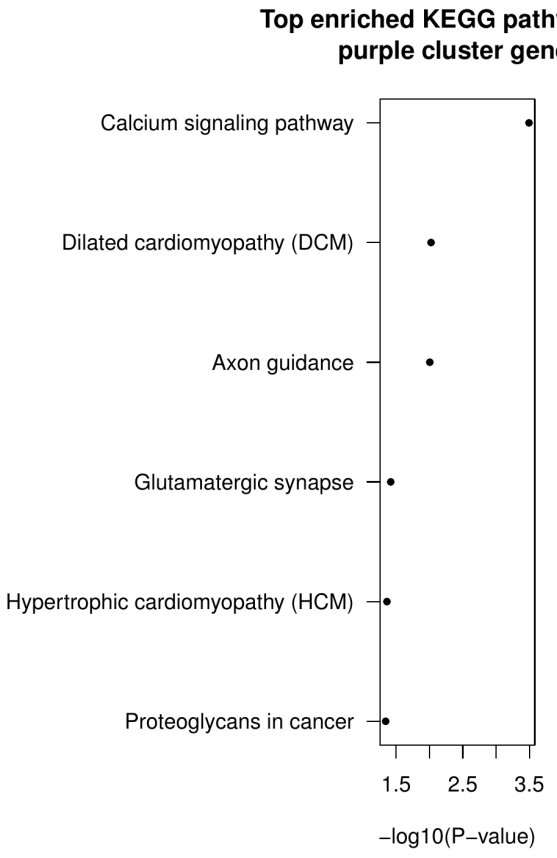
Cluster #5 Purple

The 5th cluster is the most peculiar one. The genes that are part of the purple cluster are not bound by SATB1 at their promoter sites. The genebodies seem to be occupied instead.





47,9% of the genes were not expressed.



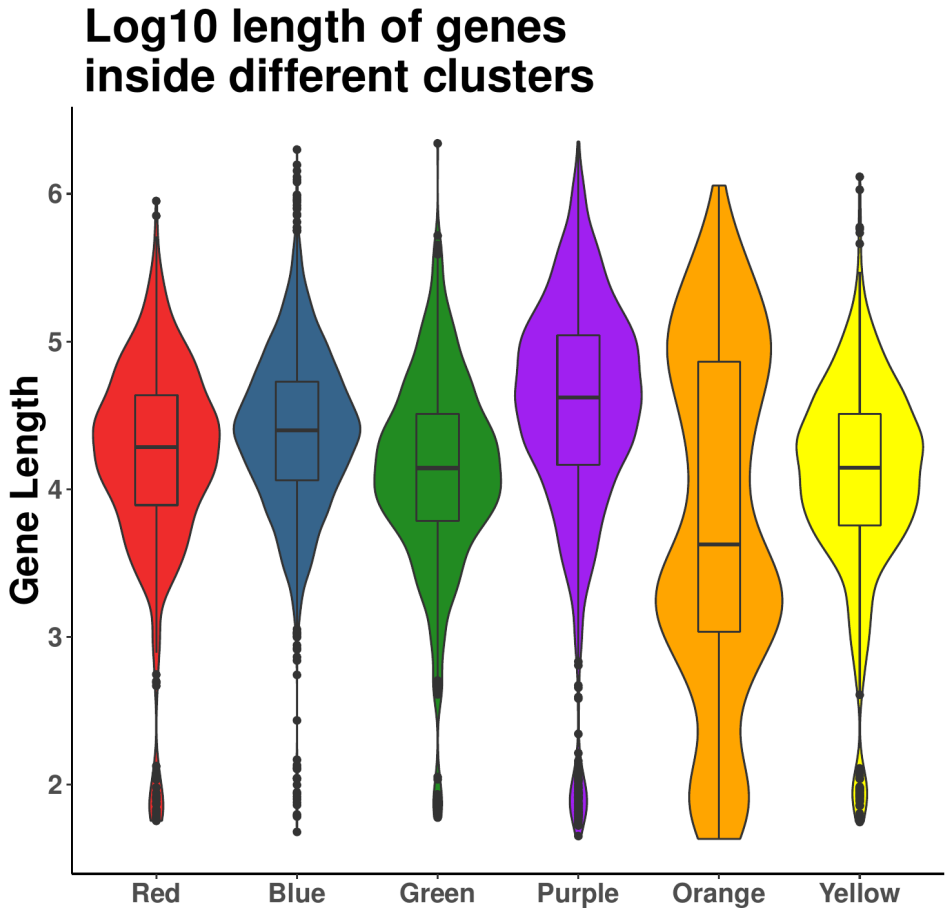
Interestingly, functional analysis of the purple cluster showed an enrichment in neuron-related pathways. The genes participating in the pathways “Axon guidance” and “Glutamatergic synapse” are the following :

Axon guidance :
16/32 genes were not expressed
Sema4f, Plcg1, Ntn4, Srgap1, Ntn1, Sema4d, Ptk2, Nfatc4, Camk2a, Pard3, Abl1, Trpc3, Sema4a, Sema3c, Sema4b, Pak1, Pak3, Nfatc3, Trpc1, Rac2, Plcg2, Sema7a, Sema3d, Trpc5, Efna5, Lrrc4c, Prkca, Epha6, Slit3, Camk2b, Ntng1, Dcc, Plxnc1

Glutamatergic synapse :
13/22 genes were not expressed
Dlgap1, Grik5, Grm3, Slc38a3, Cacna1d, Grm1, Adcy6, Grm8, Gnas, Grin2b, Adcy7, Trpc1, Cacna1a, Gng12, Gng2, Grm5, Pla2g4e, Prkca, Plcb1, Cacna1c, Grik2, Grm7, Grin2a

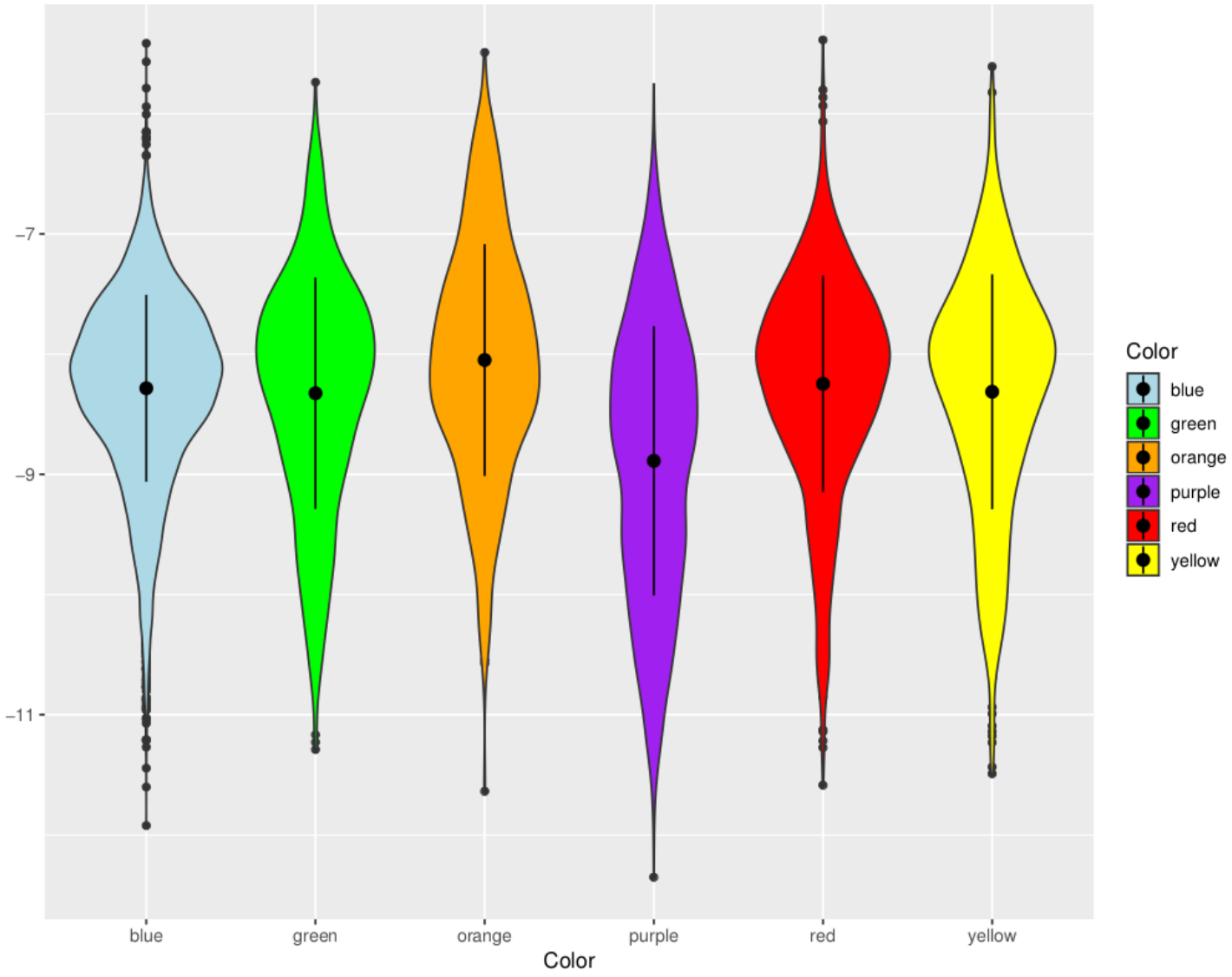
It is known that SATB1 plays a role in the central nervous system (Balamotis et al., 2011). The overall binding properties of SATB1 across clusters revealed that although SATB1 binding events occur mainly near to gene loci, SATB1 also binds a lot of genes that are silent. This binding is localized even at the transcription start site of genes, suggesting that SATB1 can bind heterochromatin.
SATB1’s pioneer factor capability was shown recently by the team of Greenleaf (Rajarshi P et al., 2018). An intriguing hypothesis is that SATB1 in thymocytes can bind a specific set of sequences, irrelevantly of whether they are in a heterochromatic or a euchromatic state, that can be relevant in another context (e.g. neurons).

The length of genes in each cluster are depicted. In general the length of the genes are pretty evenly distributed across clusters, with the exception of the orange cluster.

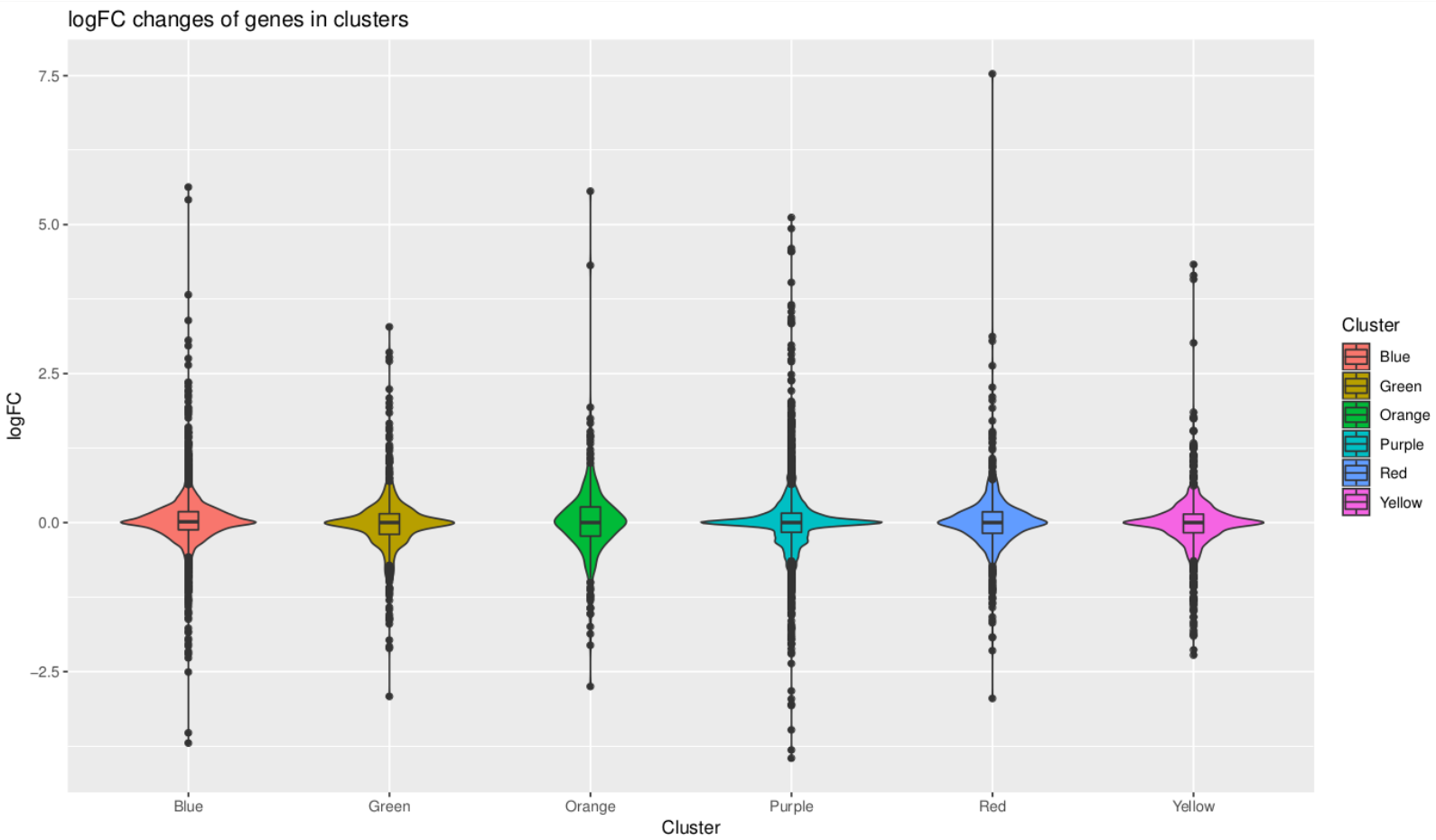


The normalized for length count values were also plotted for each cluster. The log10 transformed normalized count values were plotted for each cluster. The expression values across cluster are not different. However the purple cluster shows low expression values, in accordance with the percentage of genes not being expressed in this cluster.

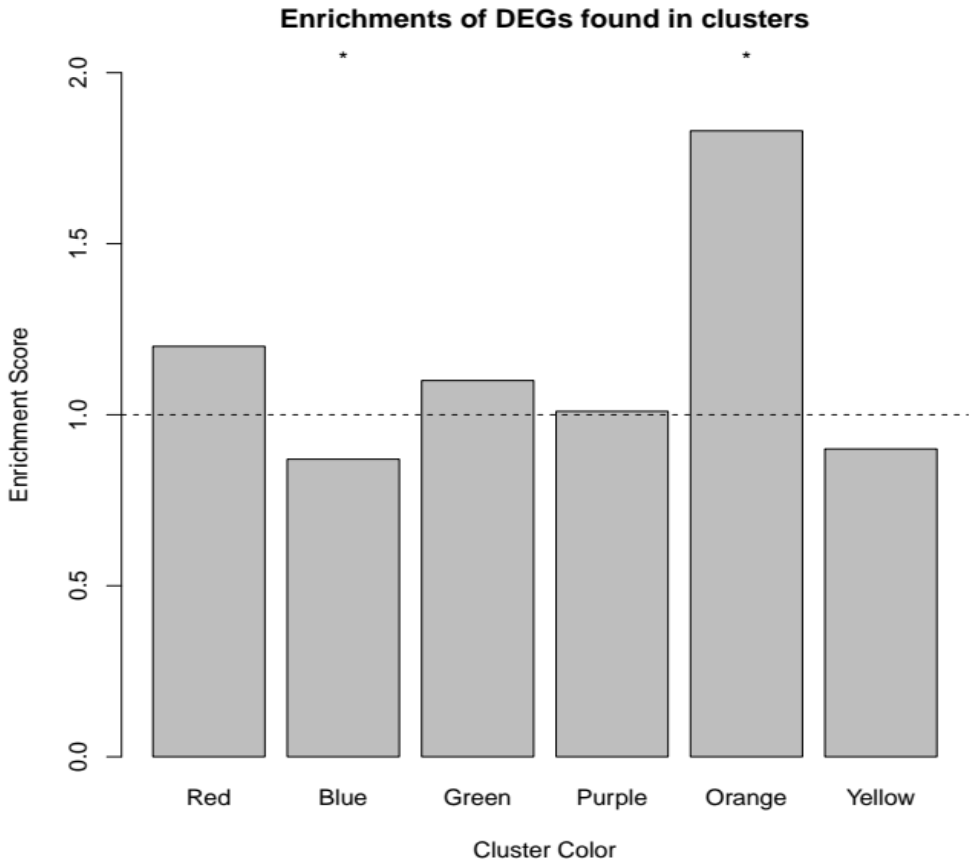
The log10 transformed normalized count values (Normalized counts per gene +0.01 divided by the length of each gene) of genes falling in each cluster were also plotted. No significant differences were observed with the exception of the purple cluster : The purple cluster contained a lot of silent genes and thus the expression values plotted are lower.



The logFC values for all of the genes in clusters were also plotted for each cluster. No striking tendencies were observed, as the number of DEGs in general is smaller than the number of genes per cluster.



Finally, enrichments of DEGs in each cluster were calculated and depicted below :



As it seems, DEGs are underrepresented inside the blue cluster. This is in accordance with the blue cluster representing ‘’housekeeping’’ functions. The genes performing such functions usually do not change expression levels. In contrast the orange cluster is statistically significantly enriched in DEGs. The orange cluster contained immune-related genes that were bound by SATB1 in multiple sites.

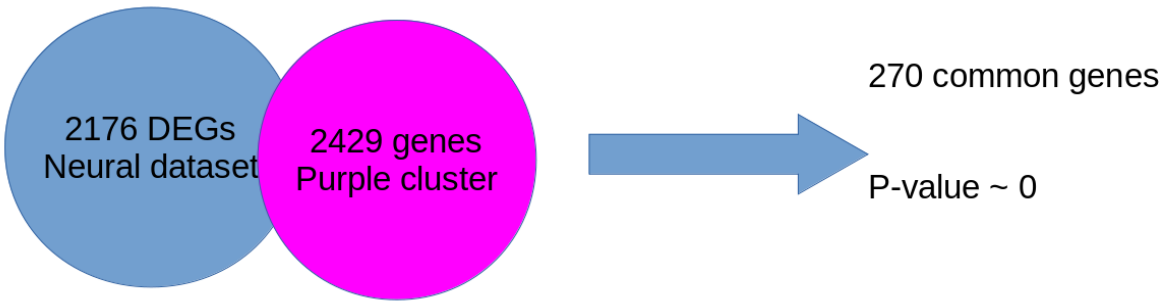
Motif analysis failed to identify a SATB1 consensus binding sequence

All the SATB1 peaks were submitted to MEME-Chip of the MEME suite (Timothy et al., 2009). No SATB1 consensus was found. Afterwards, the SATB1 binding sites for each cluster were submitted to MEME-Chip separately. Only the peaks nearby the transcription start sites of genes were taken into consideration. Once again no SATB1 consensus sequence was found.

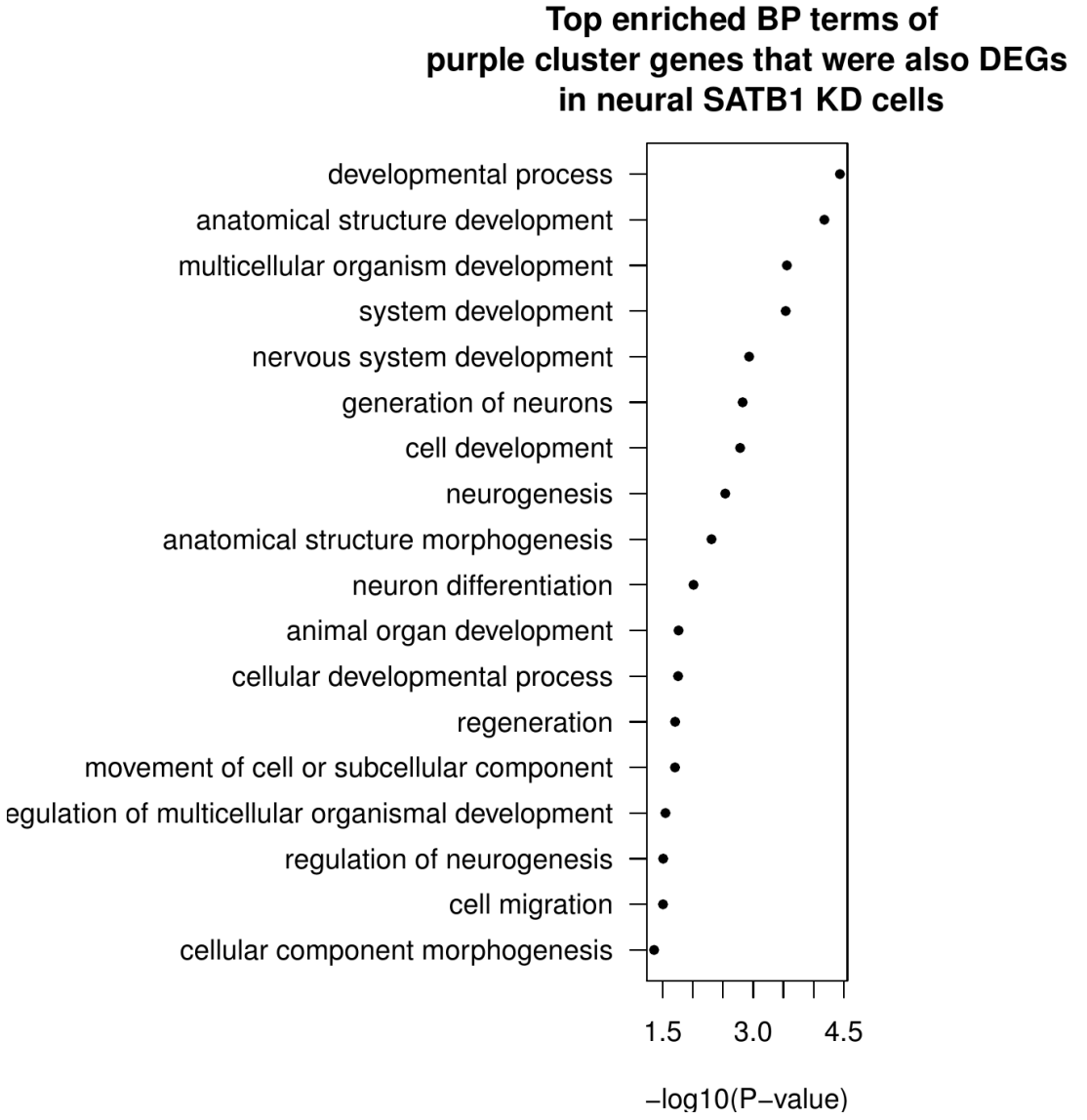
The above methodology was changed for the purple cluster, since no SATB1 binding was evident nearby the transcription start sites. For that reason, the sequences upstream of all the promoters of the purple cluster genes were retrieved (up to 1kb upstream). These sequences were then analyzed by MEME-Chip. Although a SATB1 consensus was once again not found, the motif of *Pou3f3,* a neural specific transcription factor (Sugitani et al., 2002), was found enriched in the promoter sequences of the purple cluster genes. This motif was not found enriched in the other clusters.

This finding supports the aforementioned hypothesis : Maybe SATB1 can bind specific areas in the genome of thymocytes, irrelevantly of whether they are heterochromatic, that may be under the regulation of SATB1 in neurons. If this hypothesis is correct, then we should expect that the binding profile of SATB1 in neurons should include a subset of silent, thymocyte-related genes. Unfortunately such a Chip-seq experiment hasn’t been performed up-to date.

However an RNA-seq SATB1 knockdown experiment in neurons has been performed. If the above hypothesis is correct, then genes of the purple cluster should be expressed in neurons and should be classified as differentially expressed genes in the absence of SATB1. After analyzing the publicly available dataset, the overlap of differentially expressed genes with the genes of the purple cluster was calculated. A permutation analysis was used in order to test whether this intersection is statistically significant.



The overlap was deemed as significant. Functional analysis was performed on the common subset of genes.



It seems that genes bound by SATB1 in thymocytes are deemed as differentially expressed genes in the neuronal SATB1 knockdown cells. A Chip-seq for SATB1 in neurons would shed light to this phenomenon : If thymocyte-specific genes are bound by SATB1 in neuronal populations, then it could seem that SATB1 binds a specific subset of genes independently of the chromatin landscape of these loci.

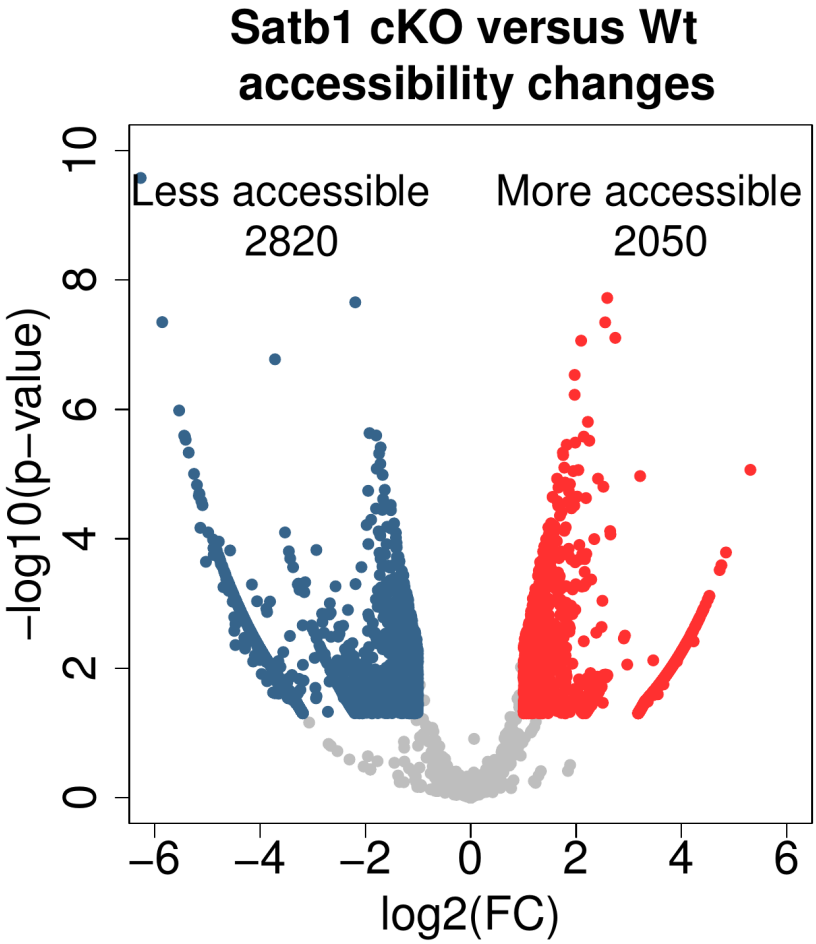
Loss of SATB1 leads to less accessible chromatin overall

ATAC-seq experiments were performed in order to detect changes in chromatin accessibility between *Satb1* knockout thymocytes and Wt thymocytes. The quality metrics for the ATAC-seq sample were excellent (see “Materials and Methods”). In order to find out differentially accessible regions, a custom pipeline, including the use of edgeR, was used. Briefly, the genome was binned in 1000 base pair bins

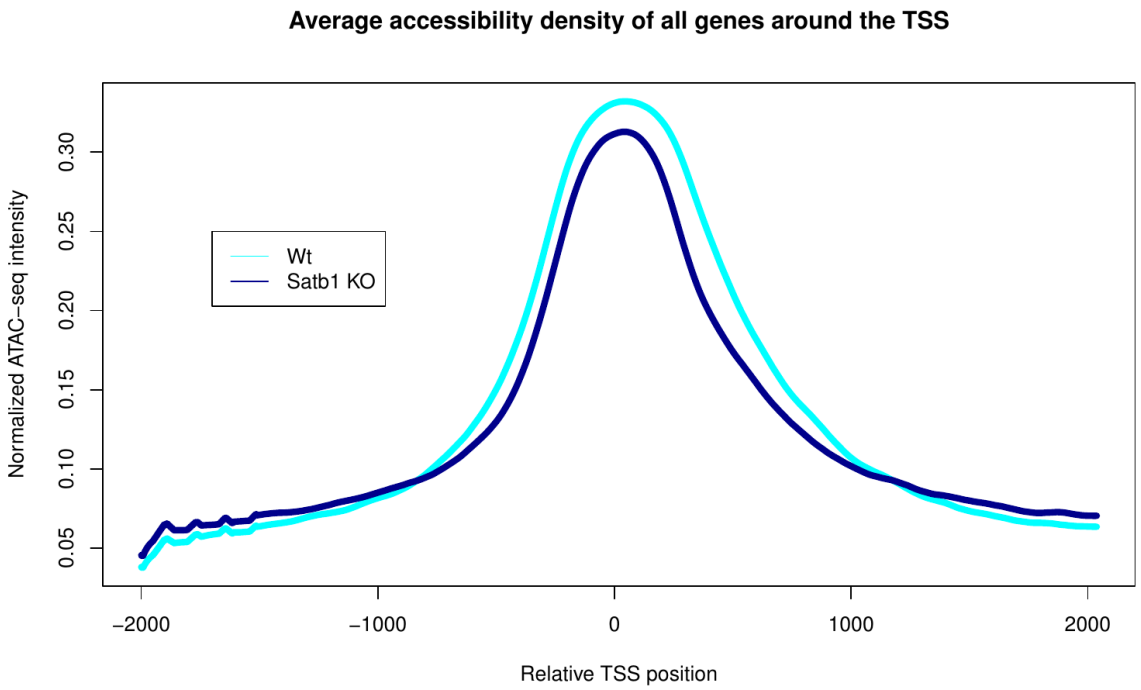
and each bin was tested for its accessibility levels between conditions. After the identification of ‘’differentially accessible regions” adjacent differentially accessible bins were merged.

The final results were the following:

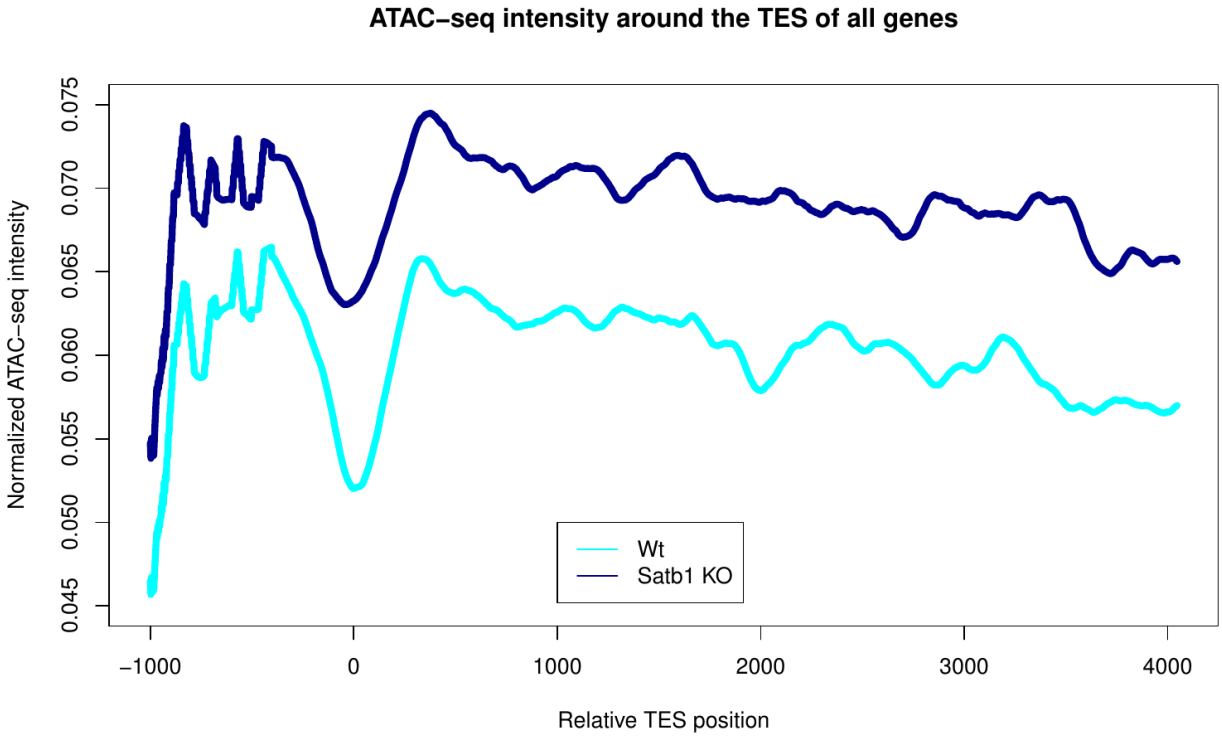
2820 **merged** bins, **less accessible chromatin**
2050 **merged** bins, **more accessible chromatin**



More regions appear to be less accessible in the thymi of *Satb1* conditional knockout mice. In order to further capture general accessibility changes across genomic loci, various average gene plots were constructed. The average gene plots were constructed with the aid of called ‘’accessible’’ peaks.



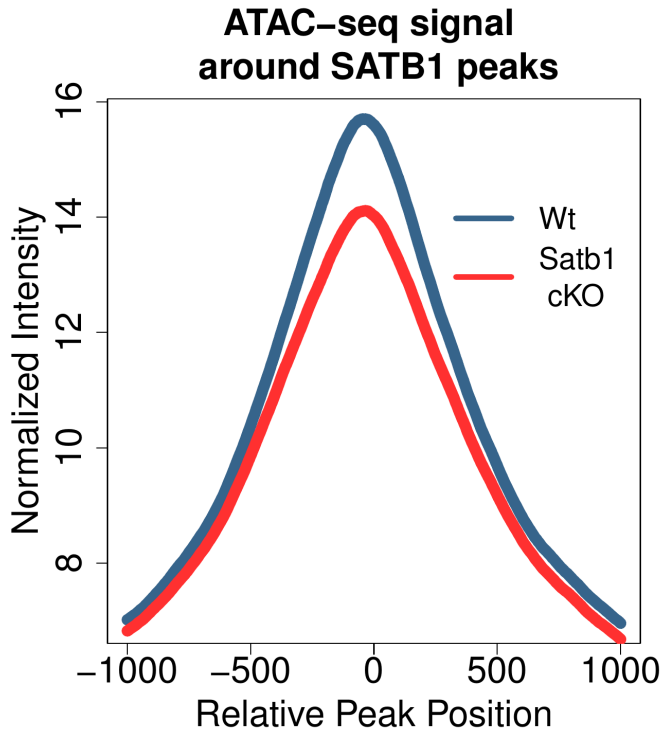
It



seems that on average, the area around the transcription start sites of genes appear to be less accessible in the absence of SATB1. An interesting finding is the exact opposite pattern for the regions nearby the transcription termination site of genes.

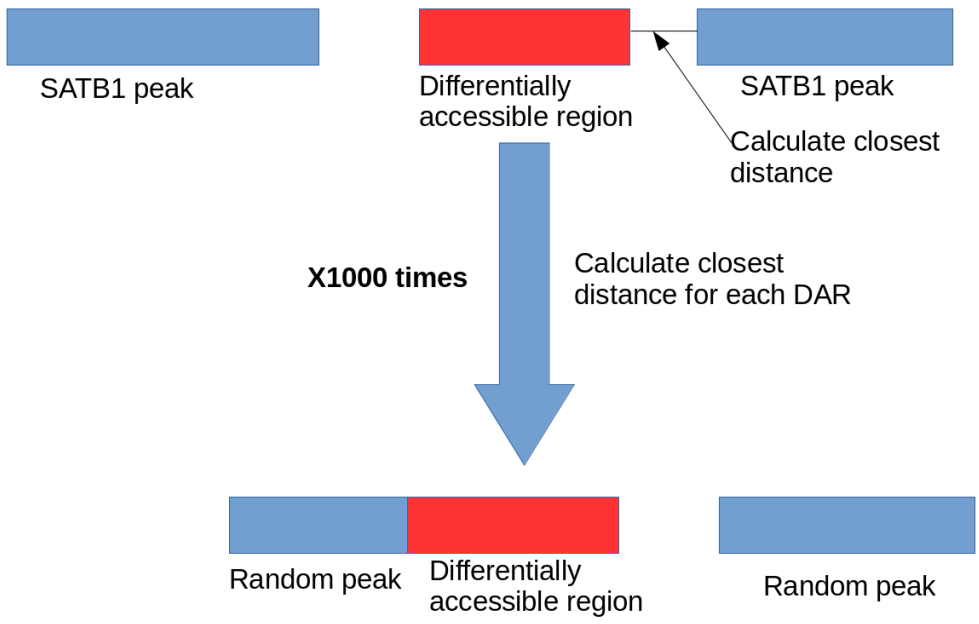
The peaks generated from the SATB1 HiChIP experiment (see “Methods and Materials”) were then used in order to test whether the areas that were originally bound by SATB1 were less or more accessible in general in the *Satb1* conditional knockout animals. It is important to note here that this dataset was used, since the ATAC-seq experiments were conducted on female mice: Krangel’s

previously used Chip-seq dataset was generated by male mice thymi (as well as the RNA-seq dataset of the sorted DP cells). The SATB1 peaks were centered and an average accessibility density was calculated around this center. The result is shown below :



It seems that SATB1 bound areas were in general less accessible in the *Satb1* conditional knockout animals. The above result however doesn't prove **causality.** It is possible that SATB1 itself isn't responsible for modulating the accessibility of the regions by recruiting chromatin re modelers or other proteins for example. Another hypothesis is that SATB1 regulates the expression of other proteins which in turn alter the accessibility of various genomic regions.

An indirect way to check about causality is depicted below :



The analysis yielded the following results :

For more accessible regions :

Mean distance to SATB1 peaks → 214641 bp
Mean **random** distance to SATB1 peaks → 54482 bp , p-value ~ 0

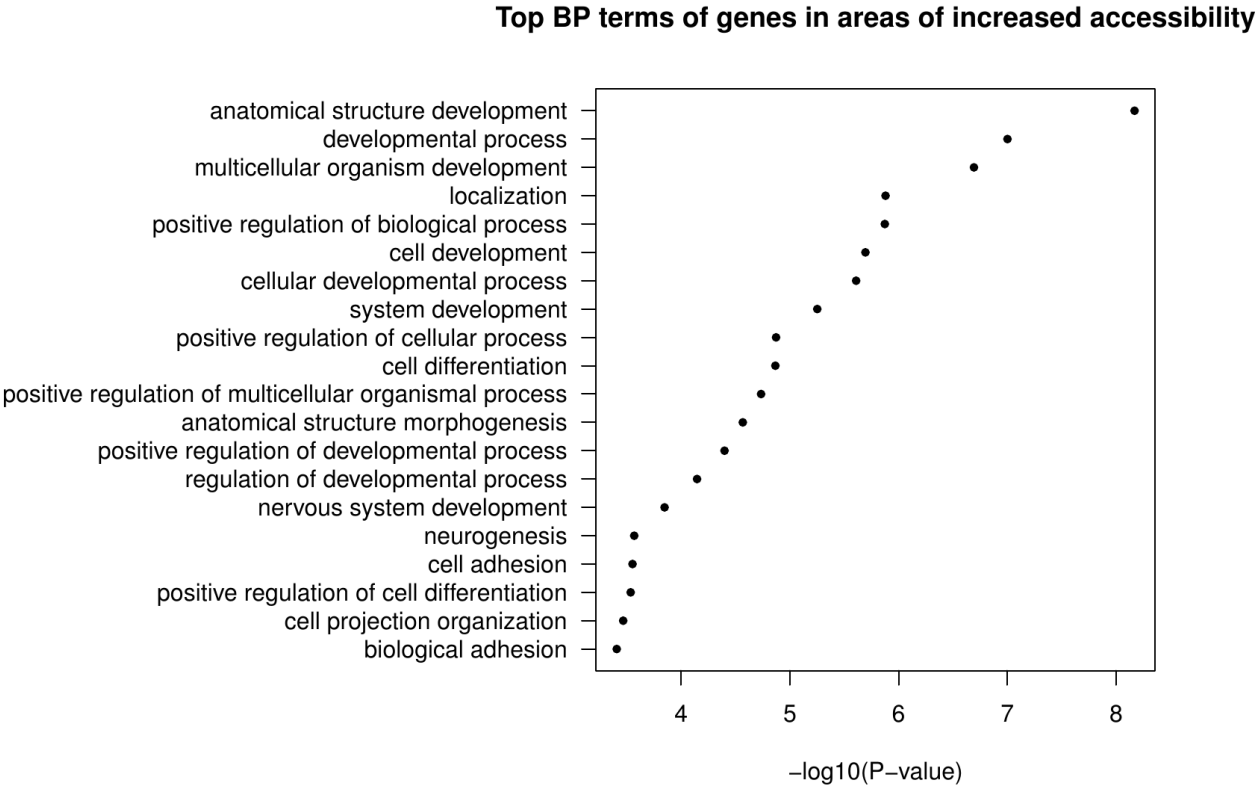
Fore less accessible regions :

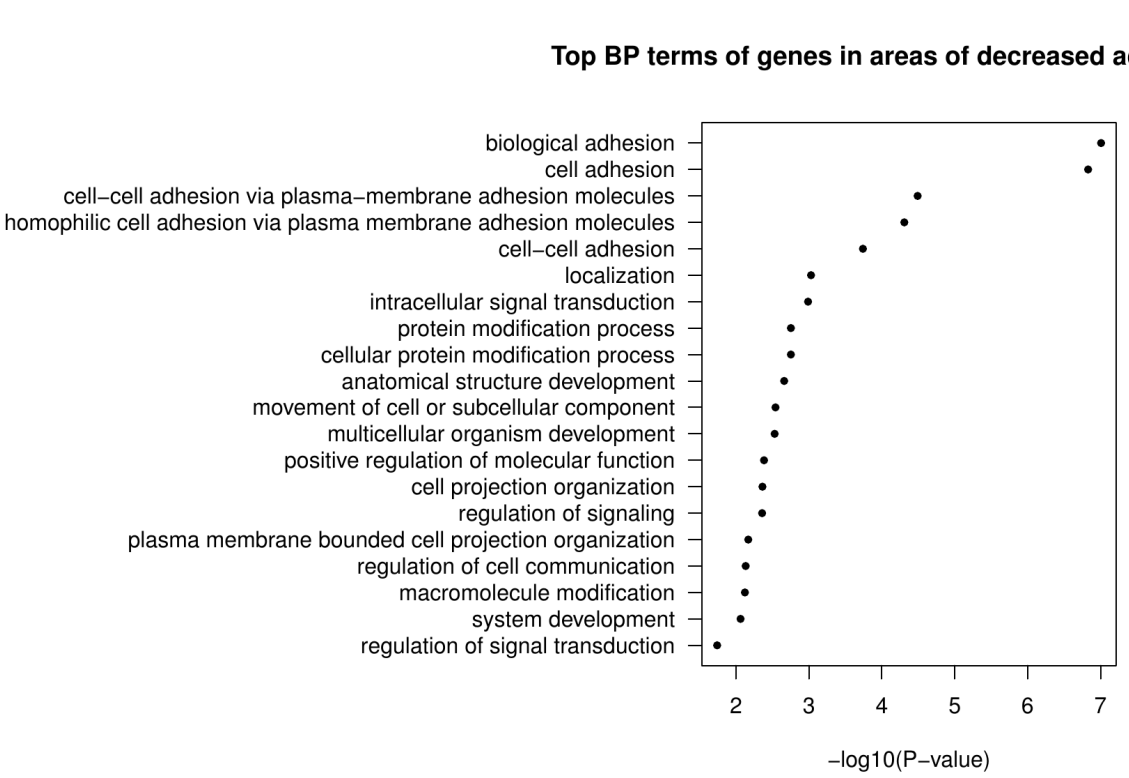
Mean distance to SATB1 peaks → 254111 bp
Mean random distance to SATB1 peaks → 54417 bp , p-value ~ 0

It seems that the SATB1 peaks are located ‘’far’’ away from differentially accessible regions. This result suggest that SATB1 isn’t directly affecting the chromatin accessibility at its target genomic regions. However another important variable is the fact that ATAC-seq was carried out on the whole thymus. Thus, changes in the accessibility of non-thymocytes (where SATB1 is not expressed) will also affect the results in the form of identified differentially accessible regions called.

ATAC-seq and stranded RNA-seq reveal a disrupted thymus environment

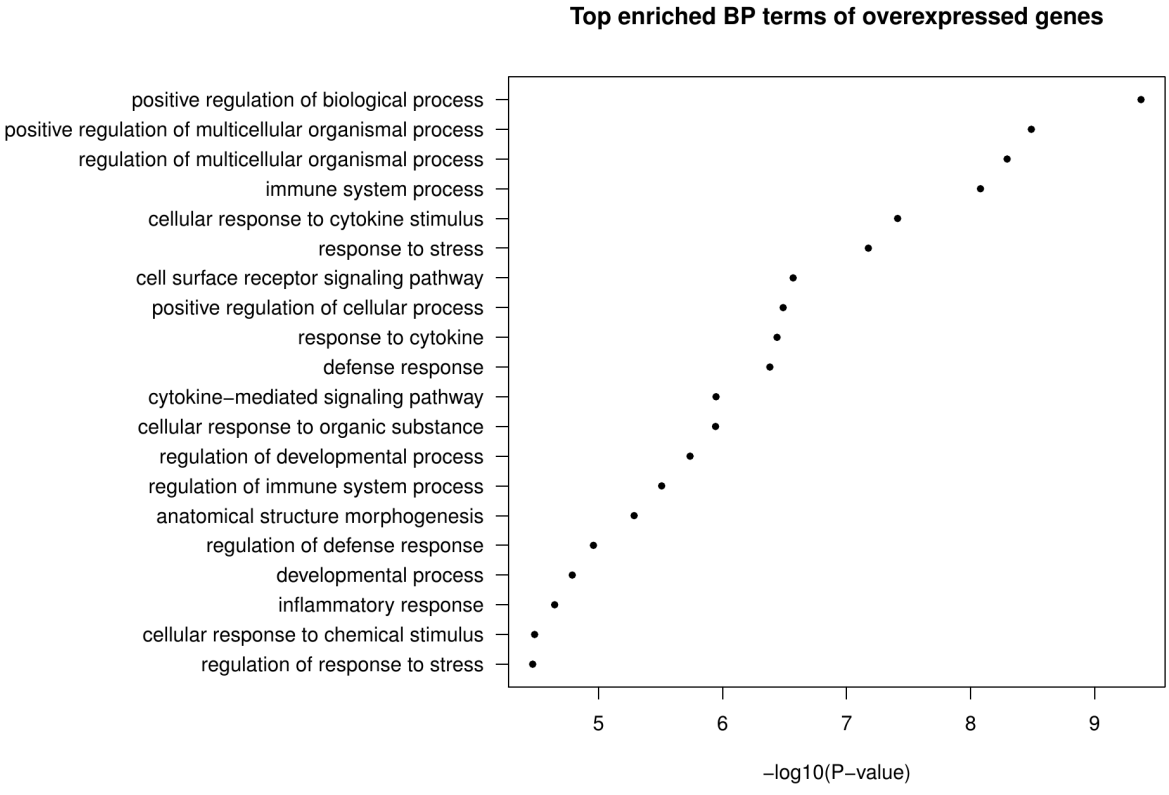
After identifying the differentially accessible regions, the gene loci falling inside them were identified. Functional analysis was performed for these groups of genes and the results are shown below :



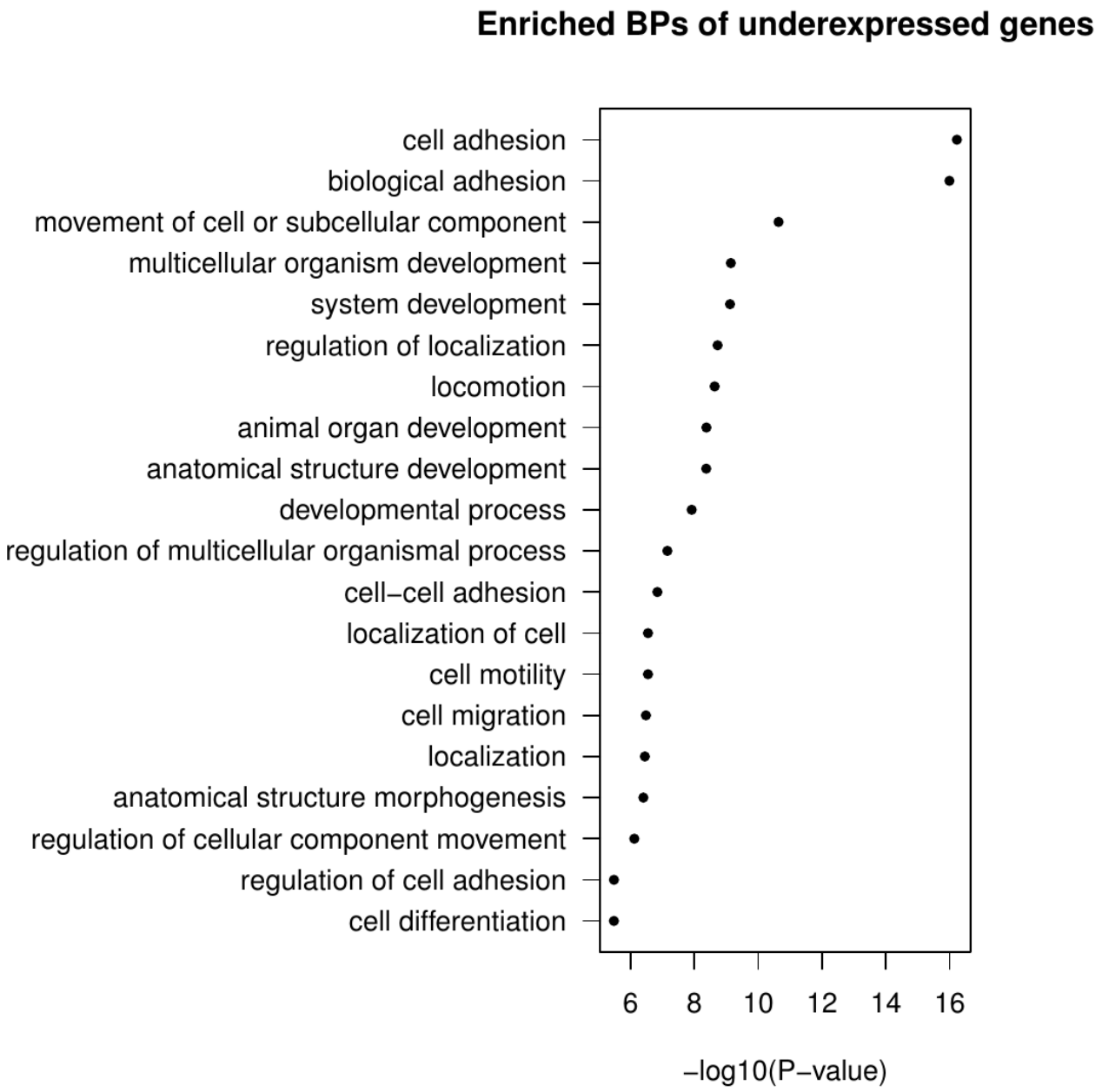


A lot of adhesion-related pathways appear to be enriched for both sets of genes. An interesting fact is that the adhesion genes found in more accessible regions are also implicated with neurogenesis. A discrepancy is evident though. The above pathways were not enriched in the differentially expressed gene subsets isolated before. A change in the accessibility of a gene locus should sometimes be reflected at its RNA levels. These changes could be attributed to the fact that ATAC-seq was conducted on female mice: It is known that the female *Satb1* conditional knockout animals develop a more severe autoimmune phenotype than their male counterparts (unpublished data from the Spilianakis lab).

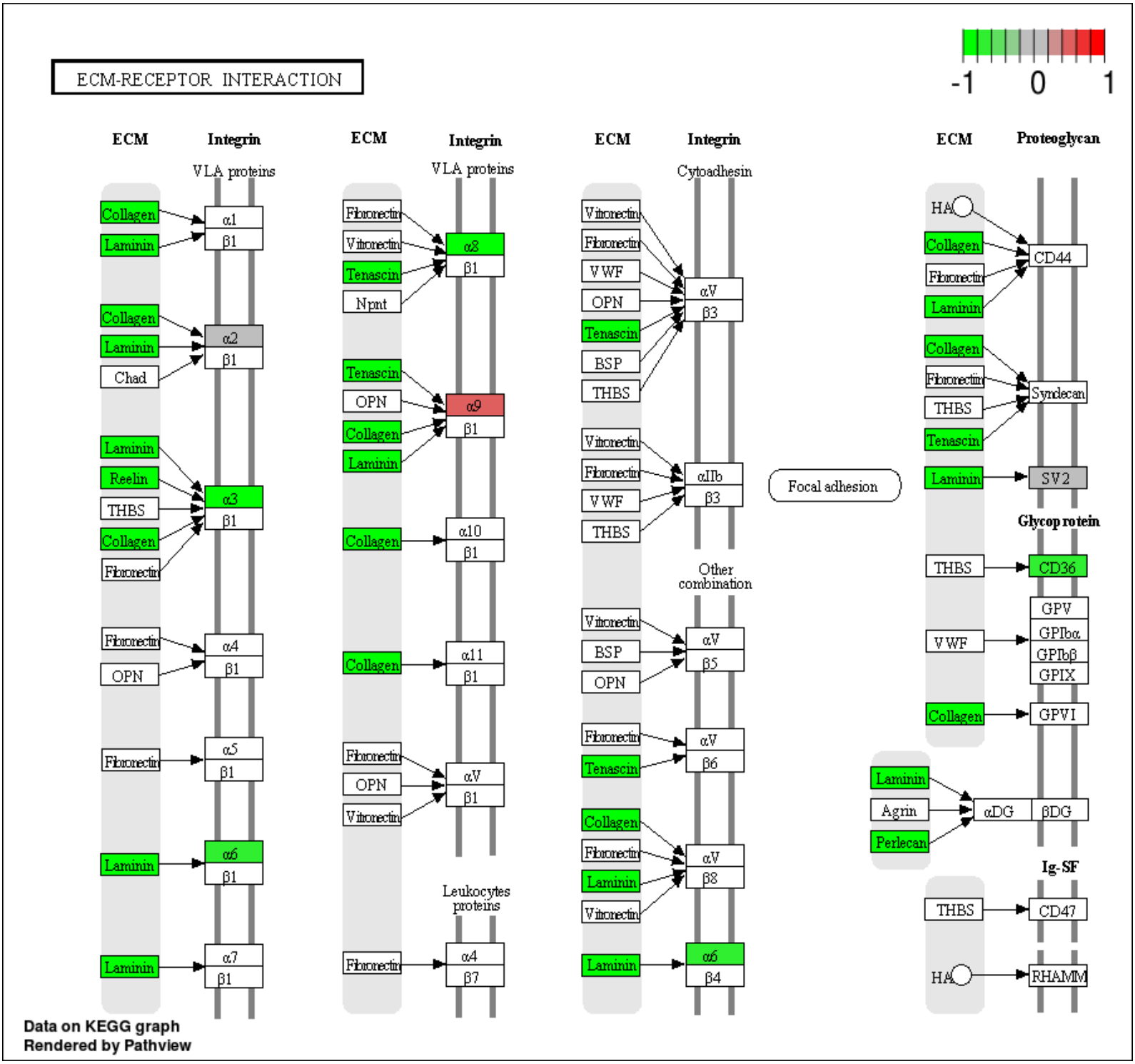
Fortunately an RNA-seq experiment in whole female thymi was also carried out. Functional analysis of the differentially expressed genes in that RNA-seq experiment shed light into the above finding.
1001 genes were deemed as overexpressed.



The large difference between the overexpressed genes of the two RNA-seq datasets is mainly due to the different annotation file used (see “Materials and Methods”) and is probably uninteresting biologically. The pathways here are immune-related (mainly associated with cytokine responses) just as the enriched pathways of the overexpressed gene of the other RNA-seq dataset.
1436 genes were deemed as underexpressed.



The pathways associated with the underexpressed genes are in accordance with the genes falling in the less accessible regions. However they are strikingly different from their corresponding counterparts.
The KEGG pathways of the above genes were also plotted :

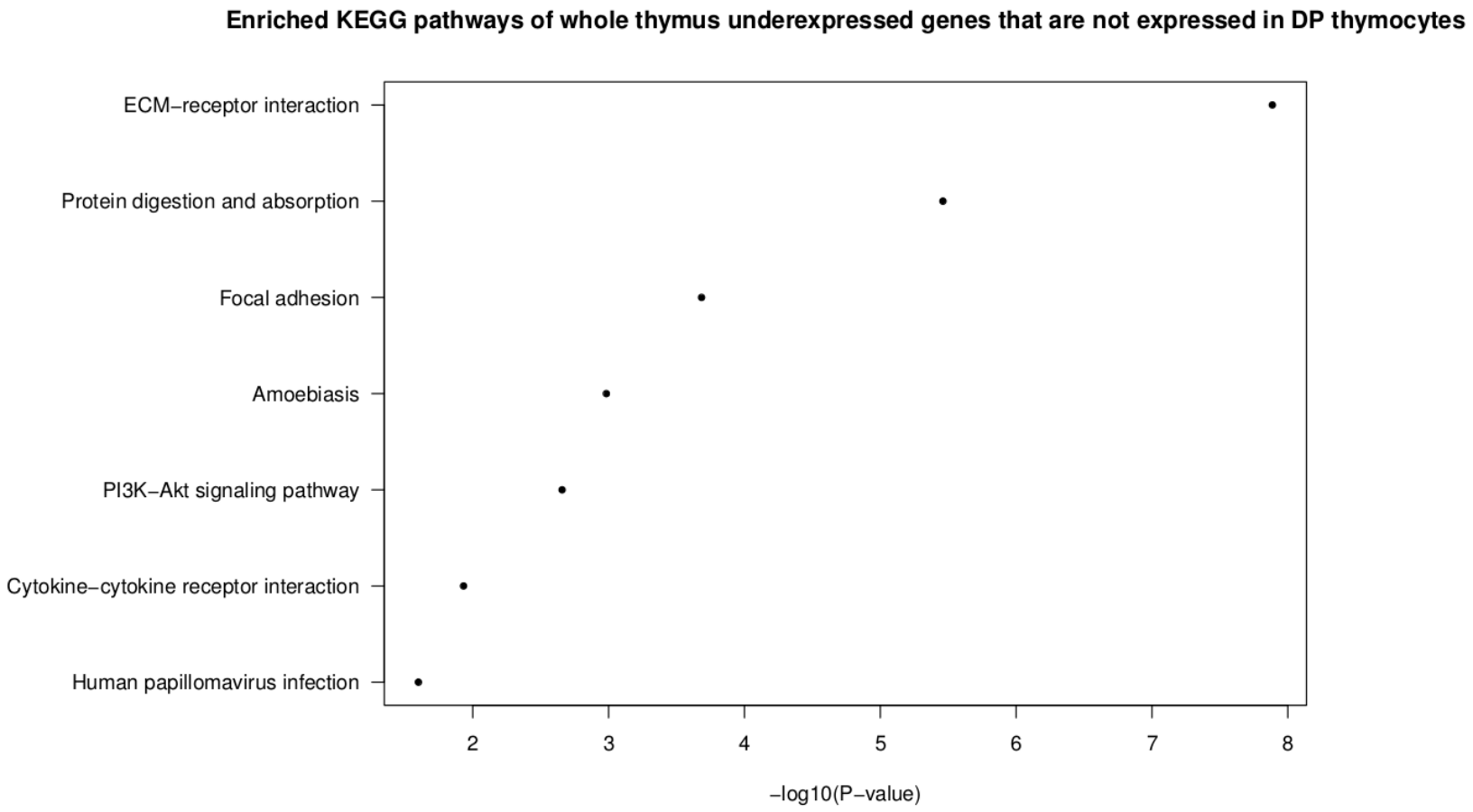


The genes of the “ECM-receptor interaction” KEGG pathway that were underexpressed or falling inside less accessible regions were plotted on the actual pathway map with the aid of an R package.

The above pathway reveals that collagen and laminin genes are underexpressed. These genes though are not expressed by thymocytes (Vieira et al., 1991). The main producers of collagen in the thymus are the thymus epithelial cells (Gameiro et al., 2010). Thus the absence of SATB1 in thymocytes somehow affects the transcriptional patterns of distinct cell populations in the thymus that do not express SATB1.

The above finding prompted us to look for more groups of genes that behave in a similar pattern: Being identified as differentially expressed in the RNA-seq experiment conducted on the whole thymus samples and being simultaneously silent in the sorted DP cells RNA-seq experiment. The following genes were isolated:

Genes being underexpressed in the thymus RNA-seq but not expressed in sorted DP cells (<=5 mean counts between conditions) : **492** genes



Genes of the “PI3K-Akt signaling pathway” :

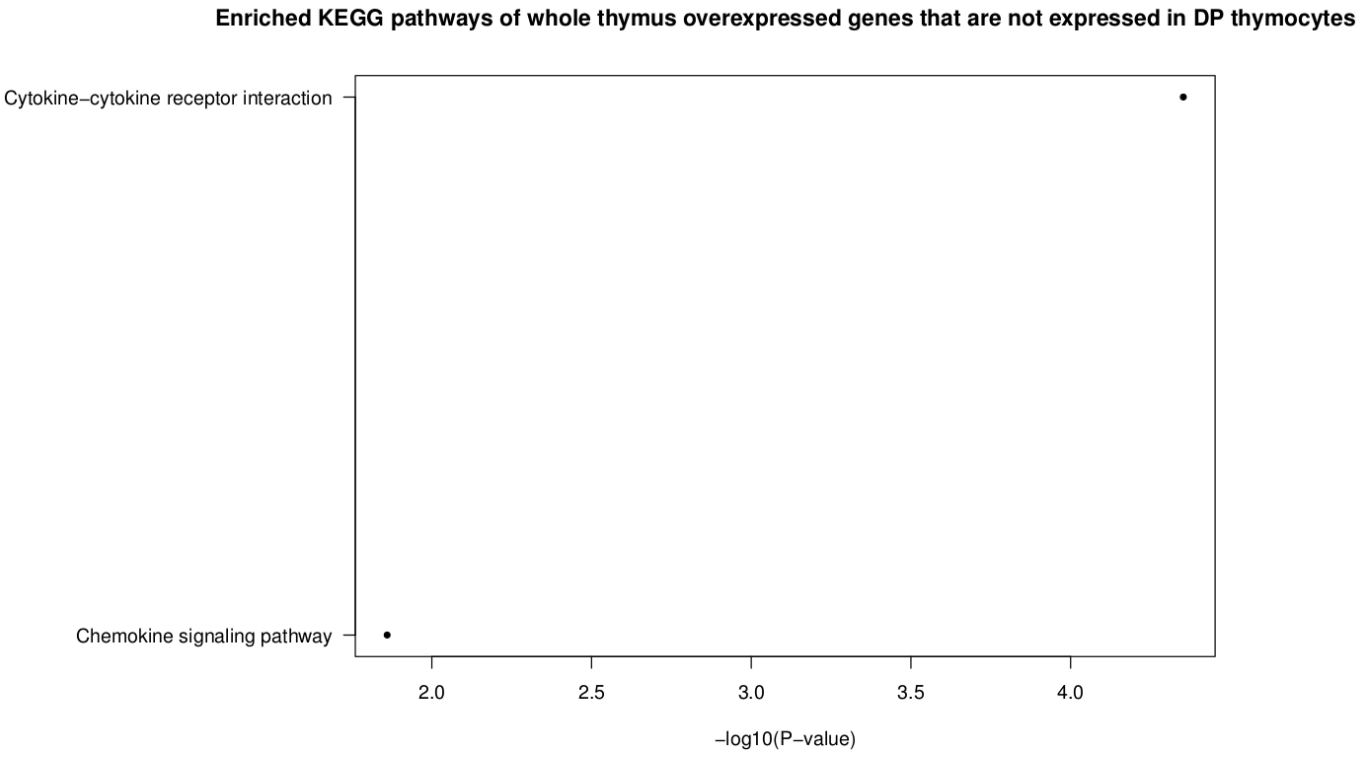
Col6a1,Col1a1,Lamb1,Prlr,Lama5,Egfr,Col6a2,Osmr,Lama3,Lamc2,Itga8,Tnc,Col1a2,Tgfa,Fgf3,Col4a5,Col4a1,Col4a2,Lpar1,Reln,Col6a6,Col6a3,Ghr

Genes of the “Cytokine-cytokine receptor interaction” :

Acvrl1,Prlr,Il13,Tnfsf11,Osmr,Il1r1,Acvr1c,Il21,Bmp3,Cd40lg,Ccl22,Ccr8,Ccr10,Ghr,Xcr1,Tnfrsf19,Cxcl12,Ifnlr1,Tnfrsf11b

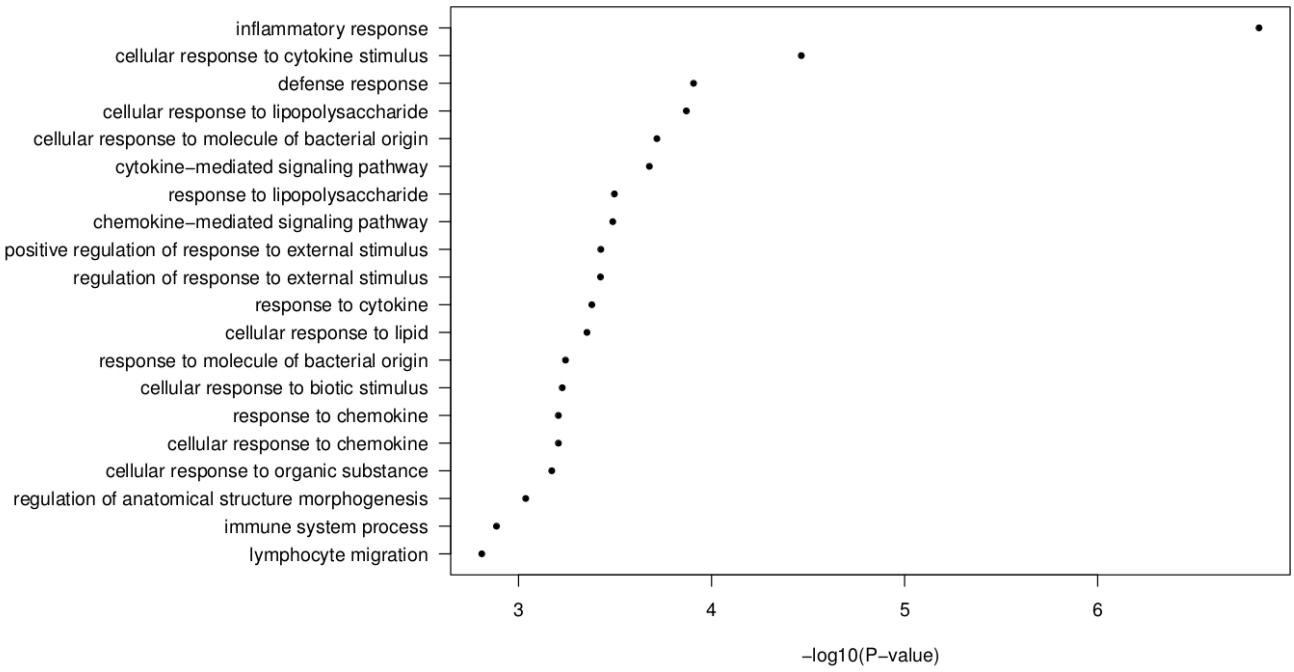
It is very interesting to note that apart from the cell-adhesion related genes, cytokine and cytokine receptors genes also appear deregulated in other cell populations of the thymus. It remains to be seen whether this deregulation is the result of an altered number of thymic epithelial cells : If the overall relative percentage of thymus epithelial cells is reduced then we should expect a drop in the expression levels of various genes that are expressed exclusively from these cells.

Genes being overexpressed in the thymus RNA-seq but not expressed in sorted DP cells (<=5 mean counts between conditions) : **175** genes



Genes falling in the “Cytokine-cytokine receptor interaction” pathway :

Ccl3,Il5ra,Ccl8,Csf1,Ccl9,Ccl1,Il1rl1,Tnfsf4,Cxcl9,Cxcl10,Ccr3,Ifng,Ccr5



Genes falling in the top biological process term “Inflammatory response” :

Ccl3,Il5ra,Ccl8,Csf1,Ttbk1,Cd5l,Ccl9,Havcr2,Ccl1,Apod,Adipoq,Casp1,Il1rl1,Cd55,Tnfsf4,Cxcl9,Plscr1,Cxcl10,Ccr3,Tgm2,Hist1h2ba,Cysltr1,C4B,Ccr5

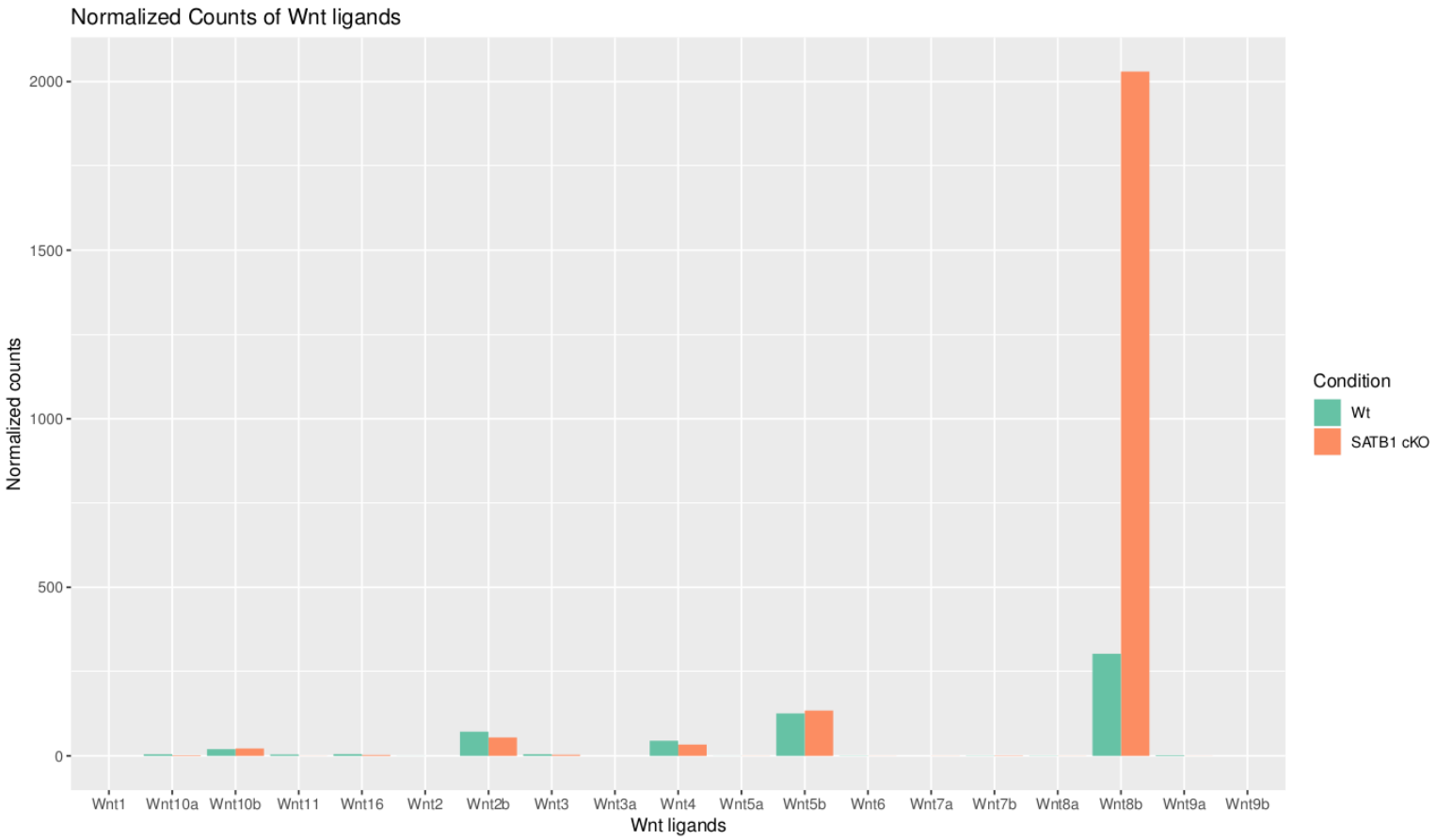
While genes related to the structure of the thymus are underexpressed, the exact opposite happens for inflammatory molecules. An example is *Ifng*, a cytokine regulating multiple types of immune responses (Savan et al., 2009).

The interplay between the SATB1 deficient thymocytes and the thymic epithelial cells seem to affect the homeostasis of the thymic epithelial cells. The thymus environment is shown to be disrupted in the *Satb1* conditional knockout mice. A molecular footprint of that disruption is evident : Adhesion related genes are underexpressed, while inflammatory molecules are produced in elevated levels by thymic epithelial cells.

There are many reasons that could lead to this disruption. A possible scenario would implicate thymocyte - secreted molecules that could affect the homeostasis of the thymic epithelial cells.

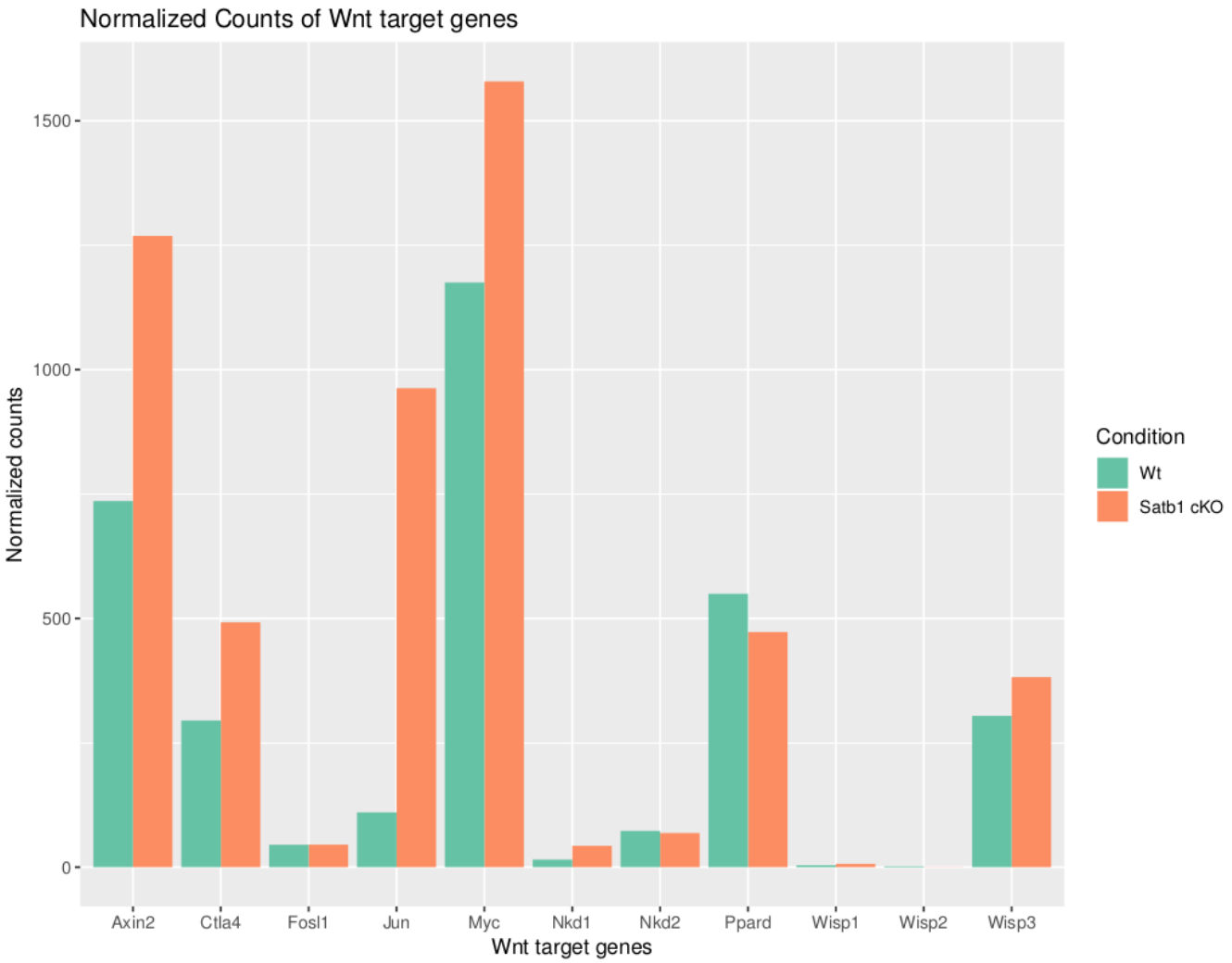
Signs of aberrant *Wnt* signaling could partially explain the disruption of the thymus environment

A lot of possible ligands could lead to differences in the homeostasis of thymic epithelial cells. However a particular ligand that is a member of the *Wnt* secretory molecules exhibited a dramatic increase in its expression levels in the absence of SATB1.



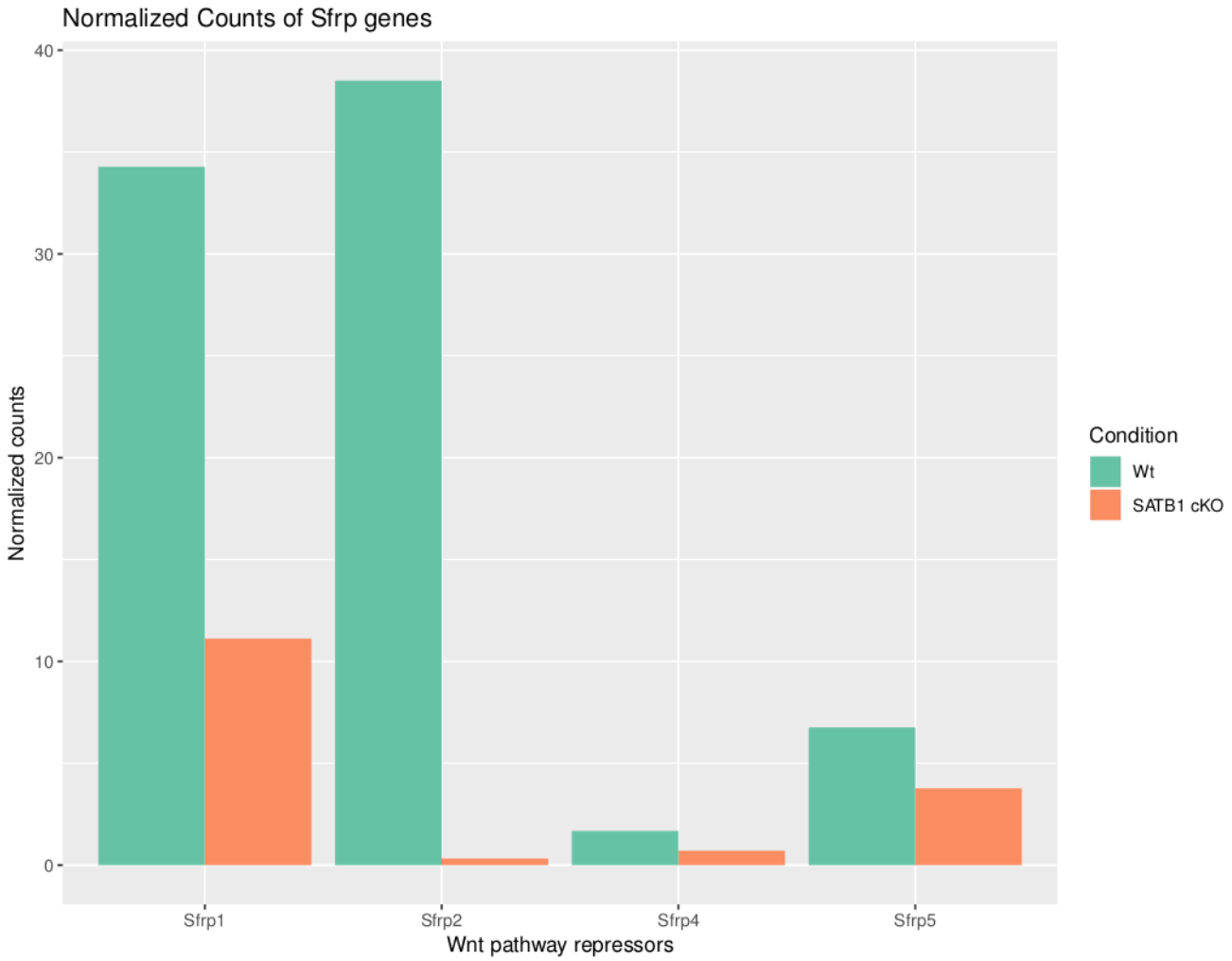
It seems that the most abundantly expressed *Wnt* ligand in the whole thymus samples, *Wnt8b*, is overexpressed in the absence of SATB1. It should be noted that its expression levels are **6 times higher** in the conditional knockout samples**.** This dramatic increase can also be observed in the RNA-seq dataset of the sorted DP cells (the increase is ~4.6 fold higher in the sorted DP cells RNA-seq), leading us to the conclusion that *Wnt8b* is overexpressed in DP thymocytes mainly.

The *Wnt* signaling pathway is known to regulate a lot of cellular processes including the adhesion of neighboring cells. If cells in the thymus are subject to elevated *Wnt* signals, then the molecular transcriptional targets of the *Wnt* signaling pathway should exhibit a rise in their expression levels. Using various sources, a subset of transcriptional targets of the *Wnt* pathway was isolated (Lustig B, et al., 2002)(He TC et al., 1998)(He TC et al., 1999)(Mann B et al., 1999)(Shah KV et al., 2008).



An overall increase across gene targets is evident. With the exception of *Nkd2* and *Ppard*, the other targets are more abundantly expressed in the *Satb1* conditional knockout. *Axin2* is a very well documented target of the *Wnt* pathway as it acts in a negative feedback loop fashion (Lustig B, et al., 2002).

The expression levels of well known *Wnt* pathway repressors were also plotted (Bryan T et al., 2009).



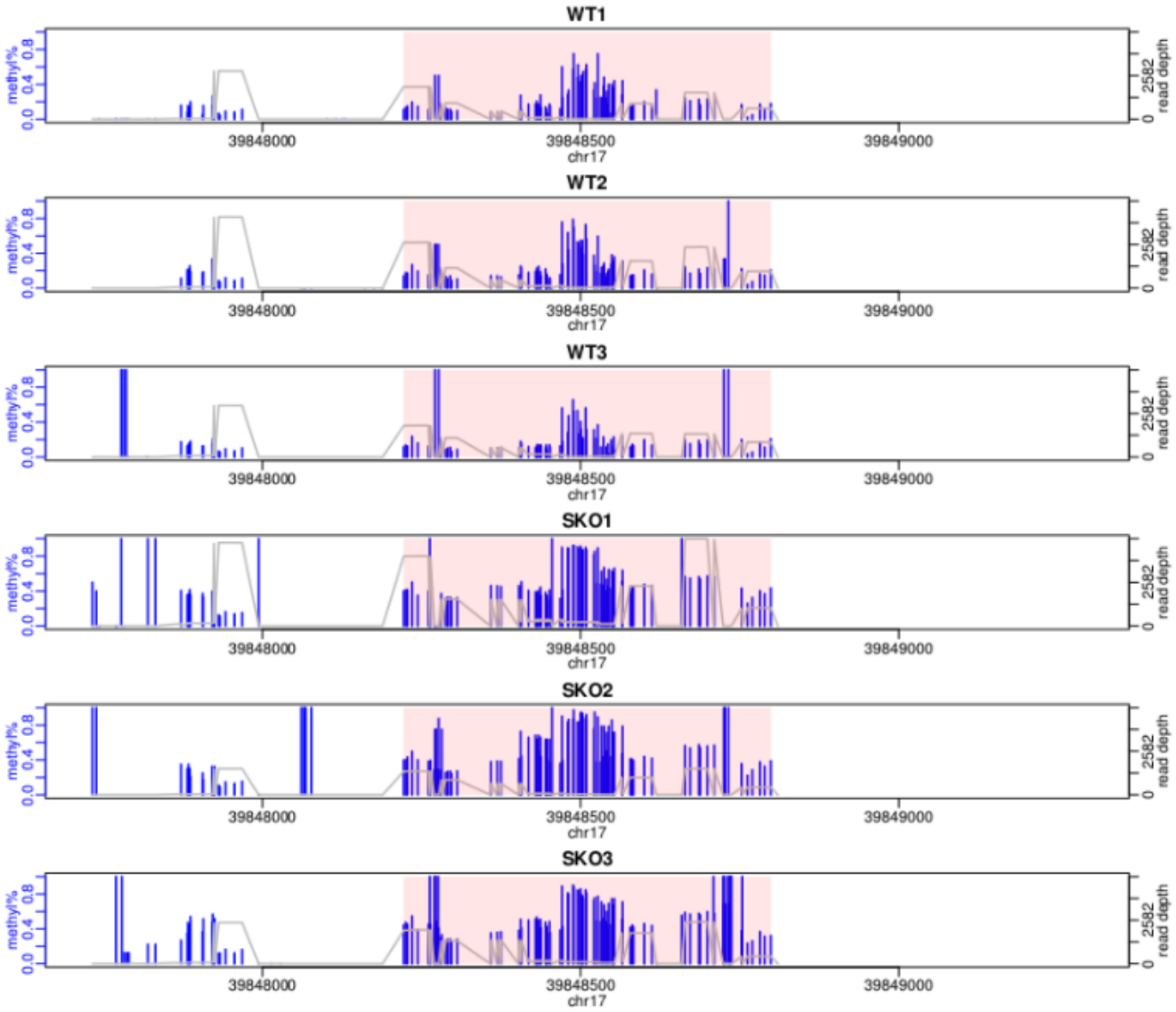
The *Sfrp* family of genes exhibit an overall decline in expression levels in the *Satb1* conditional knockout.

Overall it seems that the transcriptional level of various *Wnt* pathway associated genes suggest active *Wnt* signaling. It is currently known that the proper development and function of thymic epithelial cells is disrupted by an overactive *Wnt* pathway (Swann JB et al., 2017). Since double positive cells exist in the embryonic thymus before birth (Kuby Immunology, Sixth Edition 2006), it could be interesting to further study this possible connection between the homeostasis of the thymus environment and signaling molecules derived by malfunctioning DP cells.

Methylation levels of genes capture the disrupted thymus environment

Using a Reduced Representation Bisulfite Sequencing experiment (RRBS from now on), the methylation levels of cytosines across many genomic regions were quantified. Changes in the methylation status of genomic regions were studied in the *Satb1* conditional knockout animals versus the wild type animals. Worth noting is that the experiment was carried out in whole thymi samples.

Using an approach described in the “Materials and Methods” section, differentially methylated regions were identified. An example of a differentially methylated regions is shown below.

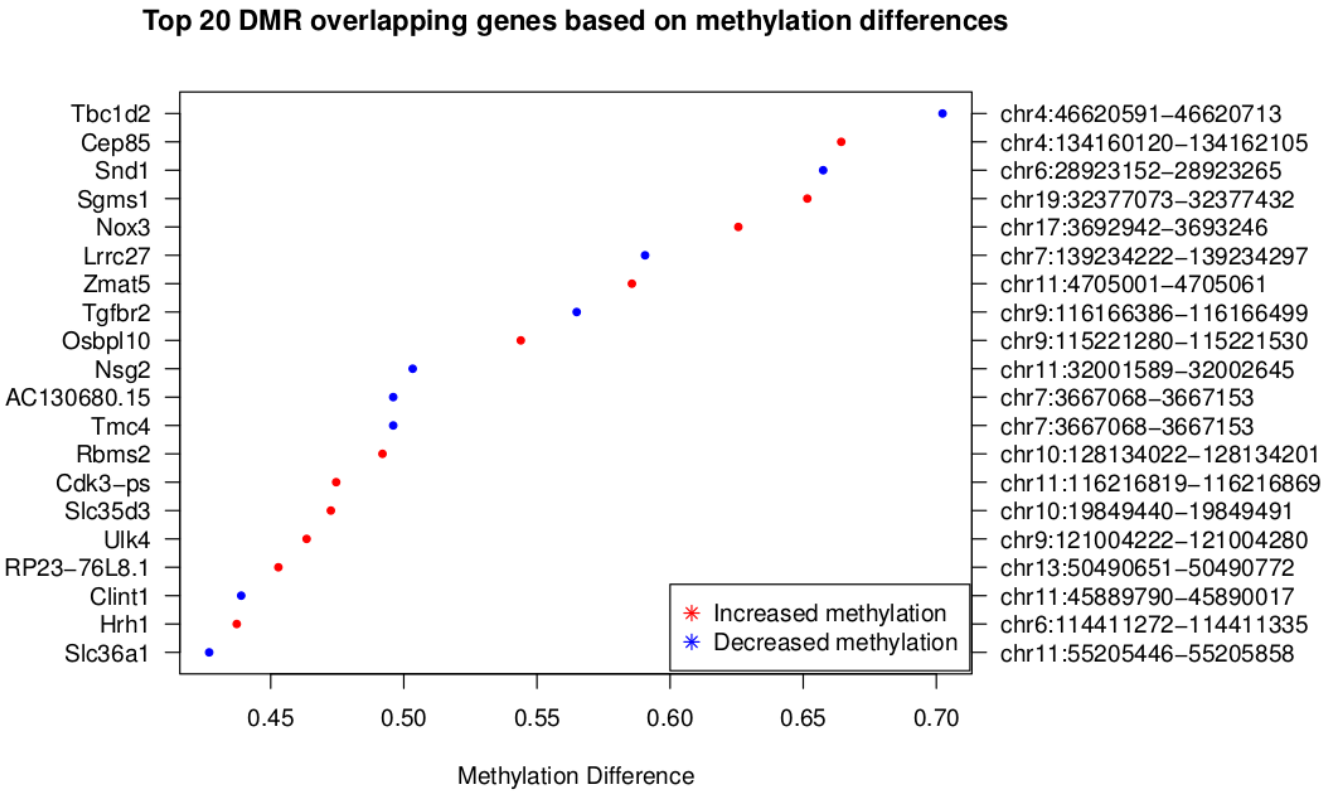


947 Differentially Methylated Regions (DMRs from now on) exhibited increased methylation levels in the *Satb1* conditional knockout.

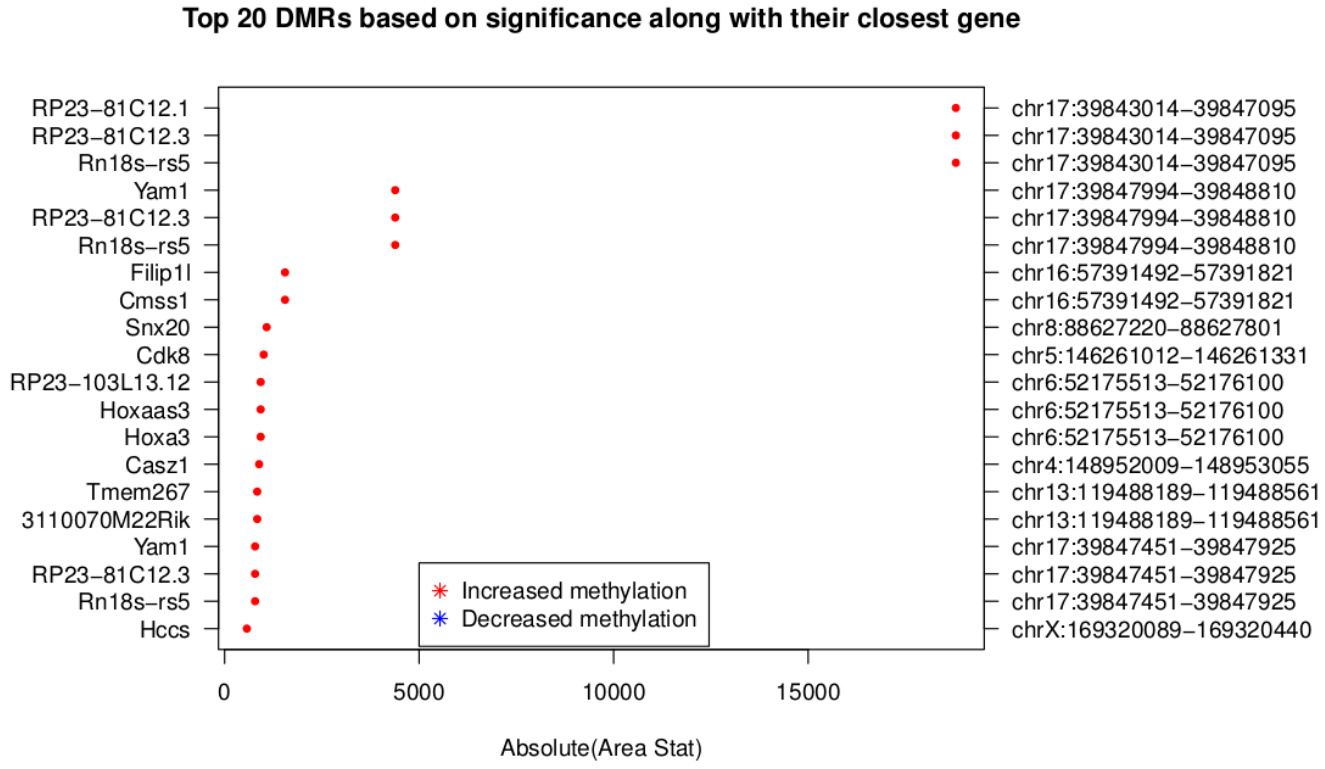
587 DMRs exhibited a decrease in methylation levels in the *Satb1* conditonal knockout.

It is already evident that the loss of *Satb1* in thymocytes leads to an increase in the methylation levels across various genomic regions. This result is in accordance with the results from the ATAC-seq dataset.

The top DMRs that exhibited the largest changes in methylation alongside their overlapping genes are depicted below:

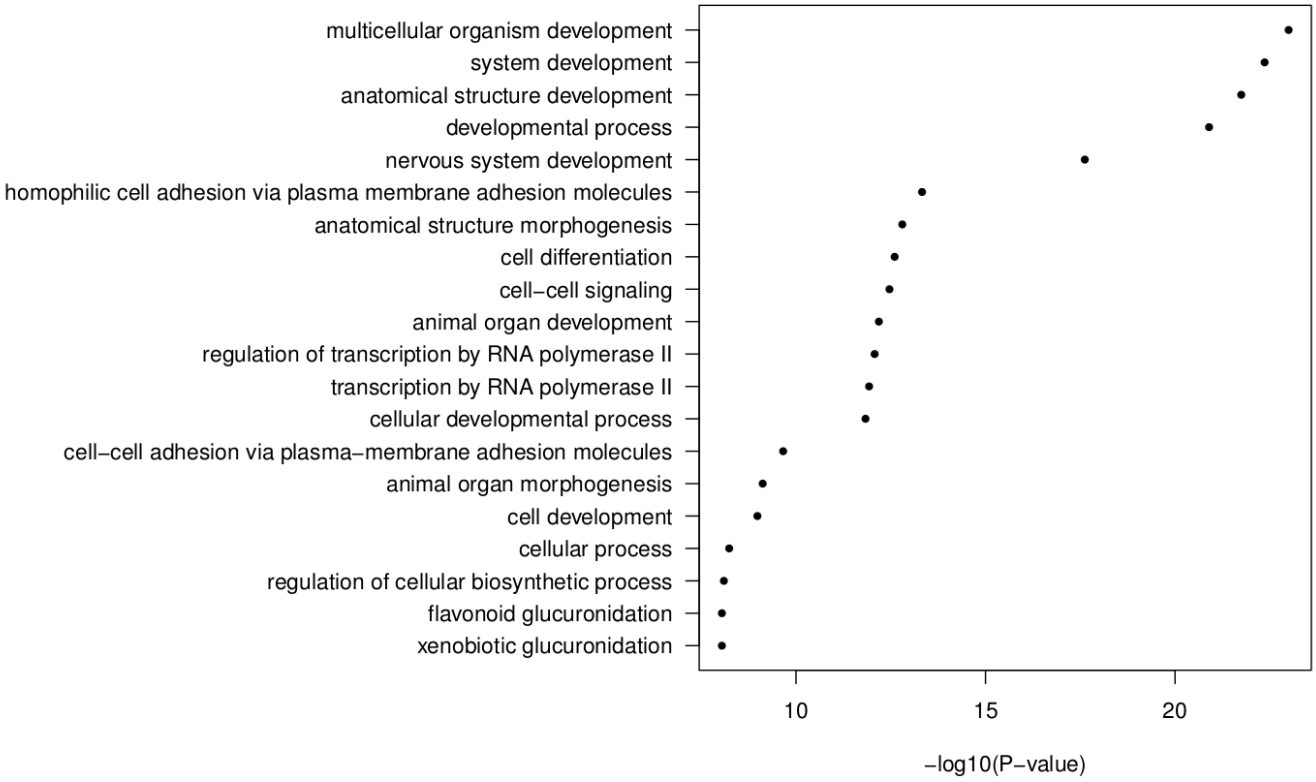


The genes that fall inside these regions are not immune cell modulators. A similar plot was also constructed by ranking the identified DMRs according to their p-values.

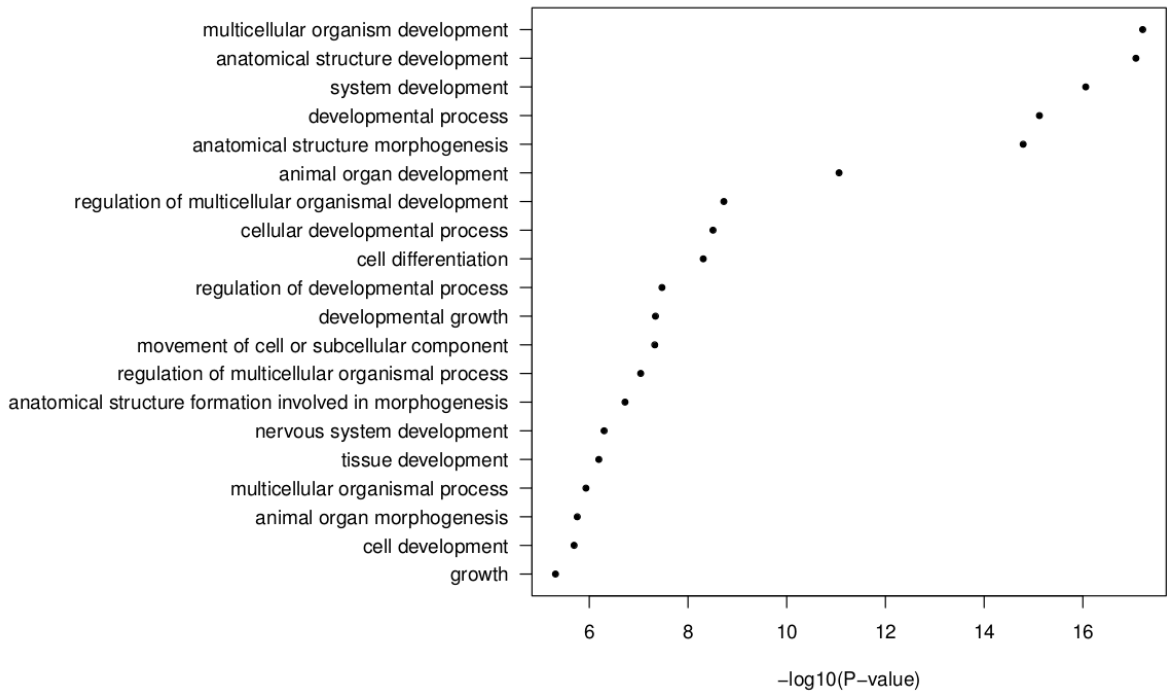


Functional analysis was carried out for the genes falling within DMRs and also close (less than one thousand bases away) to DMRs.

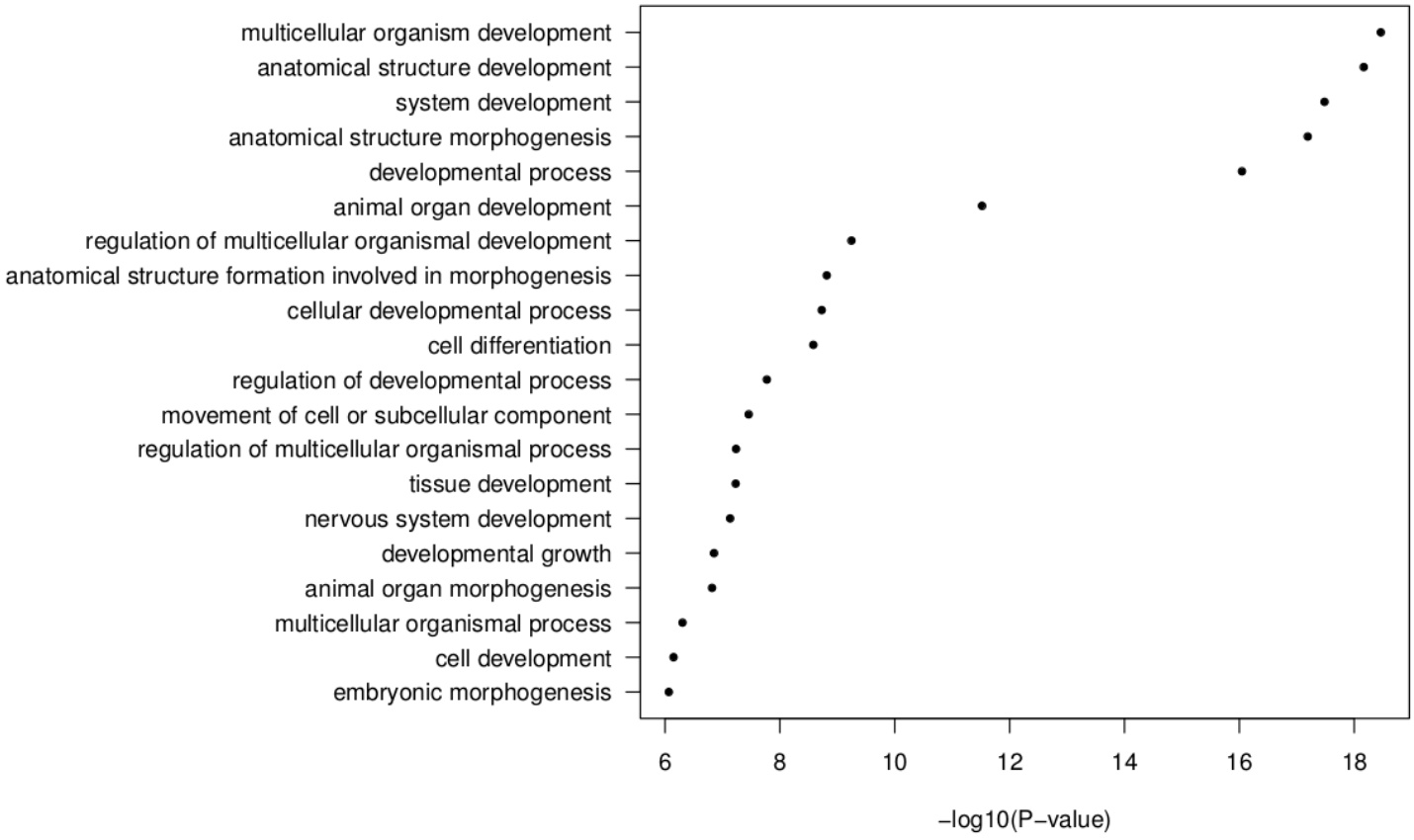
Top enriched BP terms of genes falling inside increased methylation DMRs



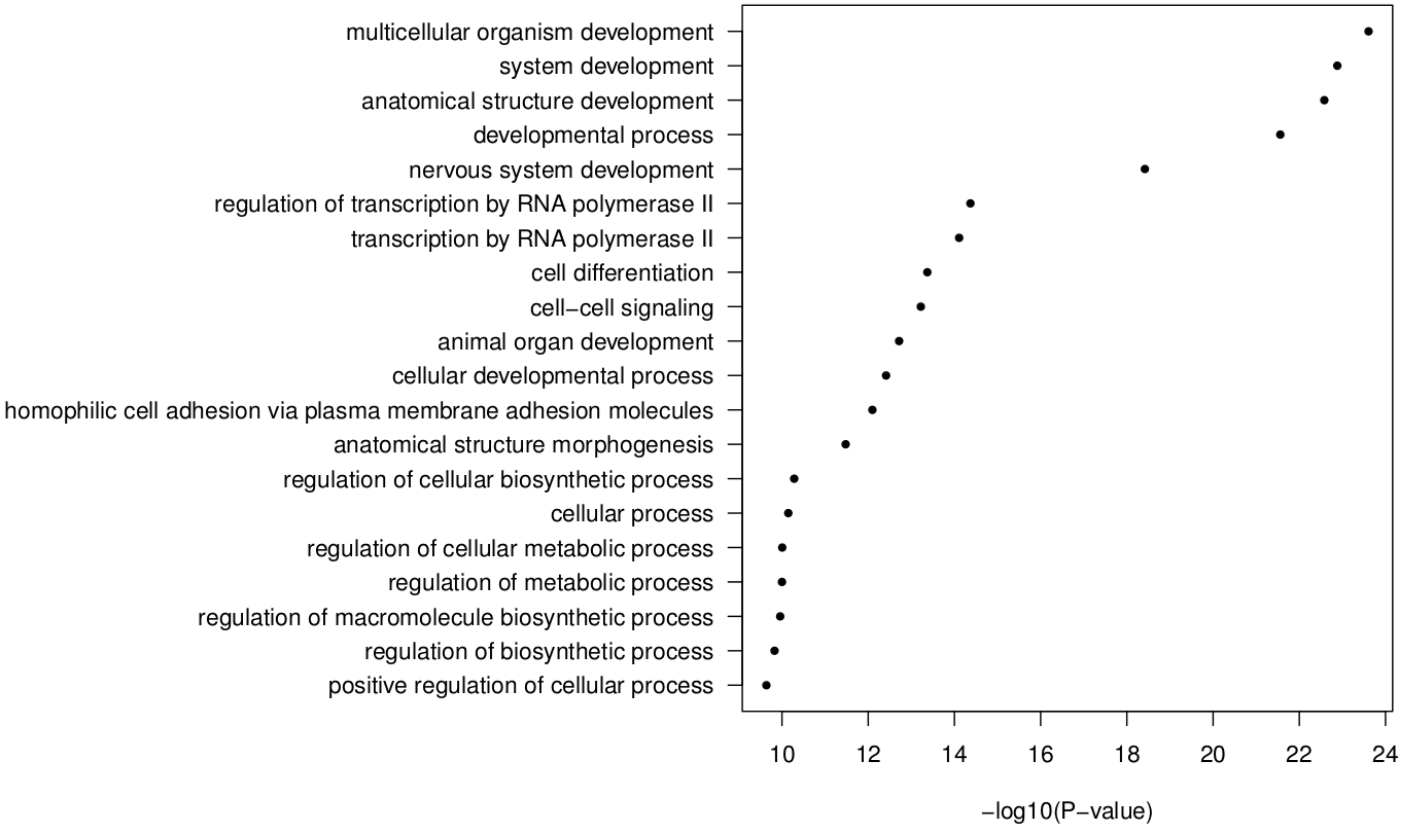
Top enriched BP terms of genes falling in decreased methylation DMRs



Top enriched BP terms of genes close to decreased methylation DMRs



Top enriched BP terms of genes close to increased methylation DMRs



Genes falling inside increased methylation DMRs appear to be related to adhesion, cell differentiation, signaling processes and developmental processes. Genes falling in decreased methylation DMRs also appear to be related to differentiation processes. A striking finding is the fact that no immune-related pathways appear to be enriched.

The next logical question is whether differentially expressed genes fall inside such regions. A table was constructed containing numeric data for differentially expressed genes falling inside DMRs. Permutation tests were carried out to check if the overlap between DEGs and DMRs is enriched.

DEGs associated with DMRs (smoothed)	**Satb1 -/- vs Wt**	**Enriched BPs/ KEGG pathways**
DEGs found inside DMRs	146/2285 p_value (bootstrap) → 0 Mean of random dist. → 63	Xenobiotic glucuronidation Tissue development Drug metabolism Pentose and glucuronate interconversions
OE genes found inside DMRs	55/907 P-value (bootstrap) → 0 Mean of random dist. → 25	Breast cancer Pathways In cancer (Wnt signaling genes)
UE genes found inside DMRs	91/1378 P-value (bootstrap) → 0 Mean of random dist. → 38	Xenobiotic glucuronidation Tissue development Drug metabolism Pentose and glucuronate interconversions
DEGs found inside increased methylation DMRs	94/2285 P-value → 0 Mean → 41	Xenobiotic glucuronidation Tissue development Drug metabolism Pentose and glucuronate interconversions (UGT cluster of genes)
DEGs found inside decreased methylation DMRs	56/2285 P-value → 0 Mean → 23	REA : MET promotes cell motility

It appears that a few differentially expressed genes fall inside such regions. Underexpressed genes falling inside DMRs are related with the term “Xenobiotic glucuronidation”. A relevant term appeared in the increased methylation functional analysis presented above (“Flavonoid glucuronidation”). Overexpressed genes that fall inside DMRs appear to be related with *Wnt* signaling components.

Another table was constructed in order to capture the behavior of over/under-expressed genes that fall inside DMRs.

OE genes found inside DMRs	Reduced/Increased 23**/33** Expected : 9/16	-/ Wnt signalling
UE genes found inside to DMRs	Reduced/Increased 33**/61** Expected : 14/25	REAC : Singaling by MET / Drug metabolism Pentose and glucuronate interconversions

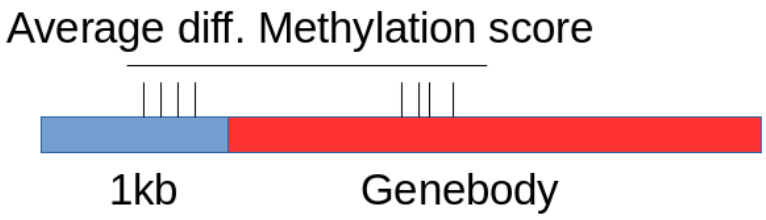
The genes do not behave as expected. Genes that are overexpressed are found in increased methylation DMRs in larger numbers than in reduced methylation DMRs. Although the majority of underexpressed genes fall inside increased methylation DMRs, a large portion fall inside reduced methylation DMRs. The above results may be a direct consequence of different cell populations in the thymus.

The underexpressed genes that were found inside DMRs and were part of the “Xenobiotic glucuronidation” were isolated :

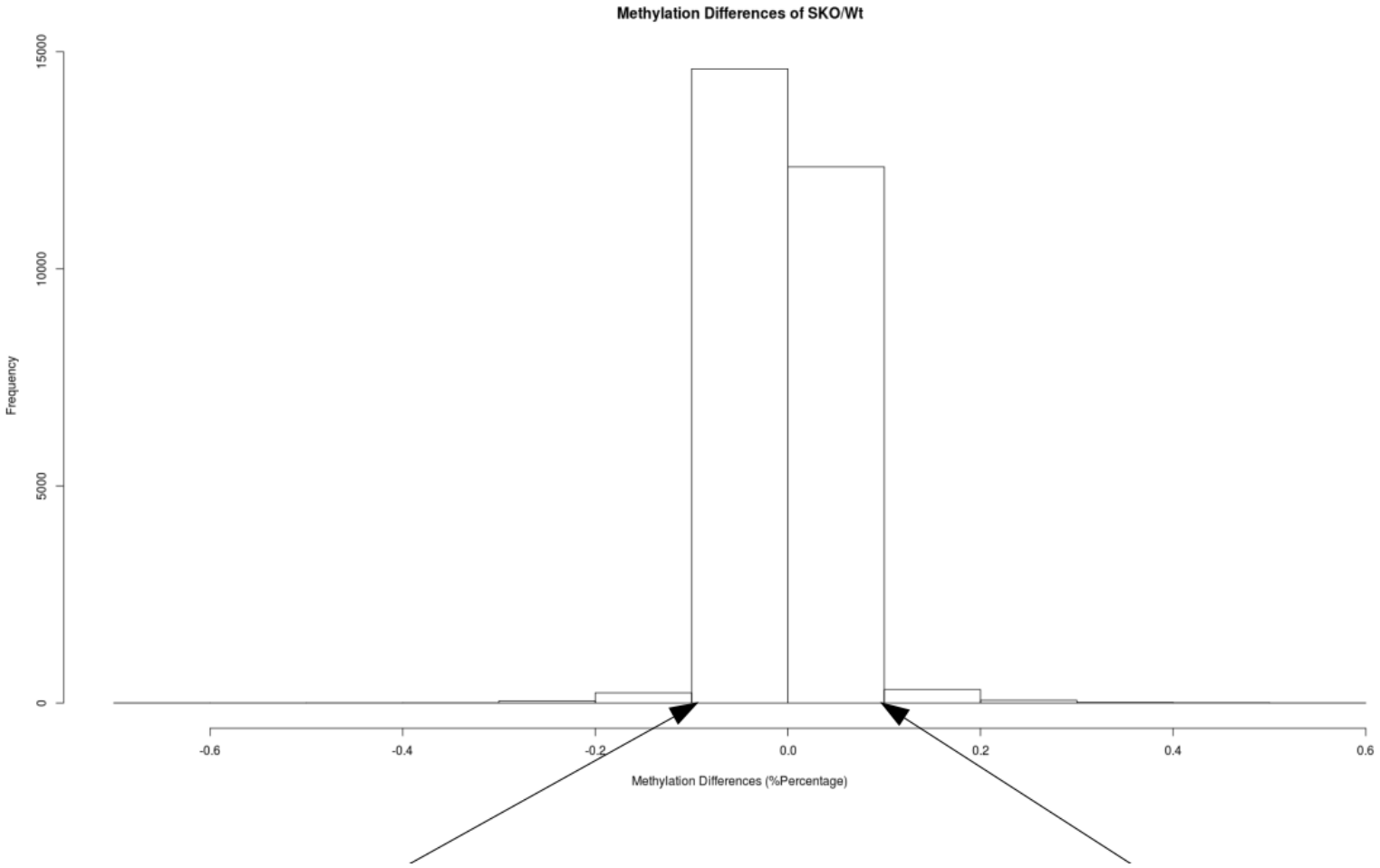
Ugt1a6a,Ugt1a5,Ugt1a7c,Ugt1a6b,Ugt1a10,Ugt1a2,Ugt1a9

Once again the *Ugt* genes were not expressed in sorted DP cells according to the RNA-seq dataset for these cells. These genes represent a case of genes being deregulated in cells other than thymocytes. The functional relevance of this pathway remains to be elucidated.

Finally a methylation score was calculated for each gene separately as depicted:



An average methylation score was calculated for each gene for both conditions. Afterwards the difference in methylation for all of the genes between the two conditions was plotted.



A cutoff of 10% difference in the methylation level of a gene was used in order to extract genes that showed large differences in their methylation status.

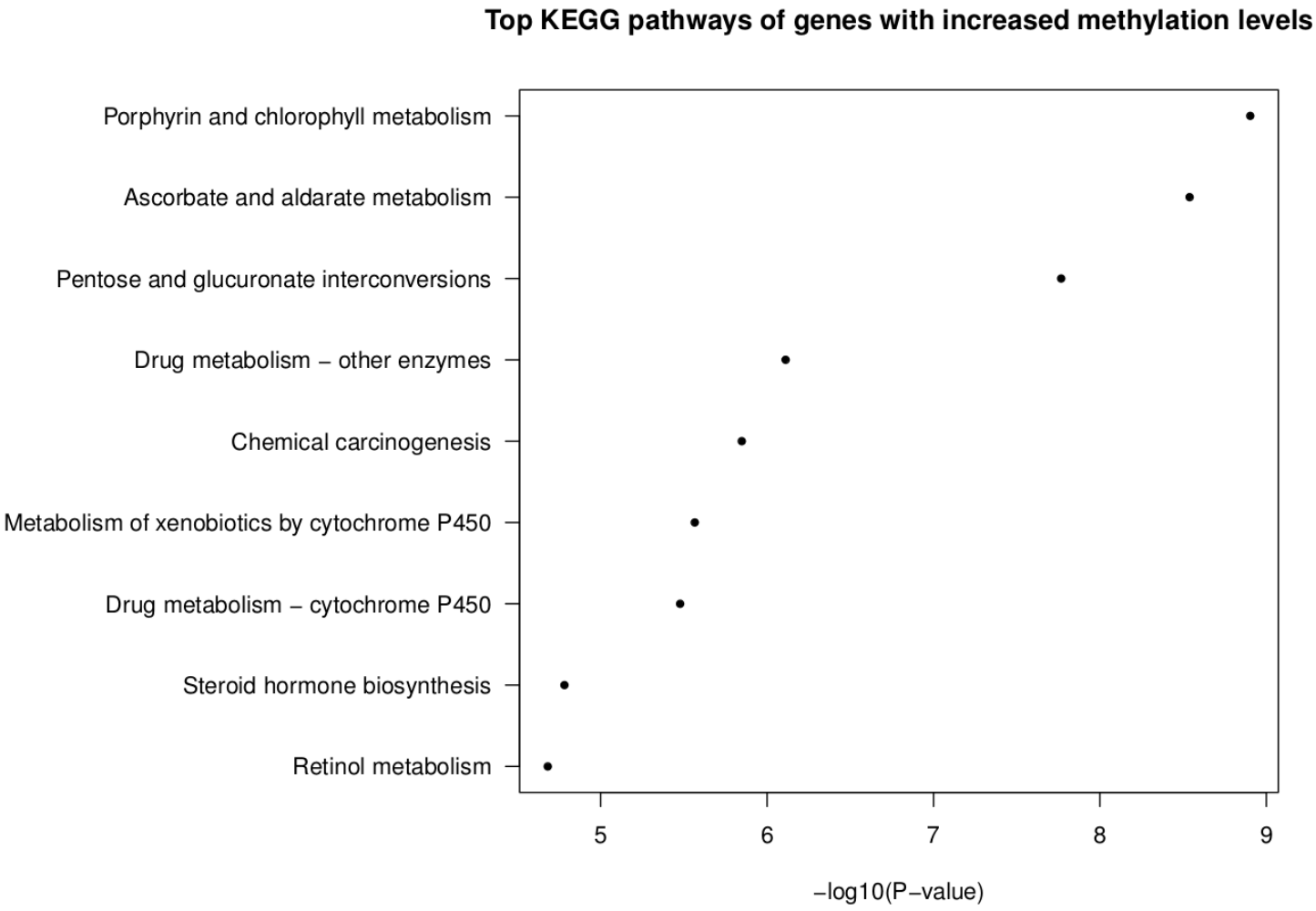
372 genes exhibited reduced methylation levels.

Just one enriched biological process term :
“Defense response” →
Pbk,C6,Pnliprp2,Klk7,Nradd,Vnn1,Ackr1,Camp,Il22ra2,Defb29,Serpinb9,Cxcr3,Cd14,Rab7b,Ilgp11,Trim5,Trim38,Ighg2c,Rpl39,H60c,Ighv8-11,Klkb1

Just one enriched kegg pathway :
‘’Arachidonic acid metabolism’’ →
Cyp2e1,Cyp2b10,Cyp2j5,Cyp2c54

The above *Cyp* family genes are not once again expressed by DP cells according to the RNA-seq experiment. The ‘’Defense’’ response pathway is the first enriched biological process related to methylation data, that somehow resembles an immune system function.

310 genes exhibited increased methylation levels.



The enriched KEGG pathways associated with the genes exhibiting increased methylation levels are practically the same across the depicted pathways. The common genes found in all of the above pathways are the following:

Ugt1a6a,Ugt1a5,Ugt1a7c,Ugt1a6b,Ugt1a10,Ugt1a2,Ugt1a9

The same genes have already been classified as underexpressed genes. The functional relevance of this gene family is something that needs more investigation.

SATB1 and CTCF are found bound in proximal sites in the genome

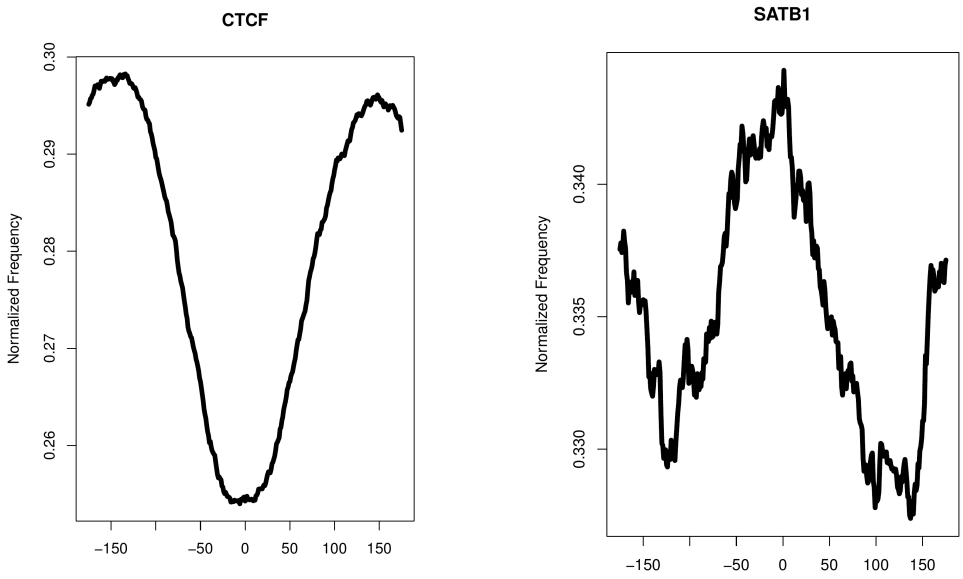
CTCF is already expressed in thymocytes prior to SATB1 accumulation. A logical hypothesis is that SATB1 affects the genomic landscape in a way supplementary to CTCF. Contrary to this hypothesis though, it is known that the proteins do not interact physically in vivo (unpublished data from the Spilianakis lab). Nevertheless they may share common binding patterns across the genome. For this reason the binding patterns of the two proteins were compared and the results were the following :

Krangel's SATB1 chip-seq experiment : 23846 peaks
9822 peaks overlap with at least one CTCF peak → Expected overlap : 1666 peaks

Hichip SATB1 peaks : 23297 peaks
4241 peaks overlap with at least one CTCF peak → Expected overlap : 529 peaks

This indicates that the two proteins may occupy sites that are in close proximity. Worth noting is the fact that **half** of the SATB1 peaks (Krangel's dataset) seem to overlap with a CTCF peak.

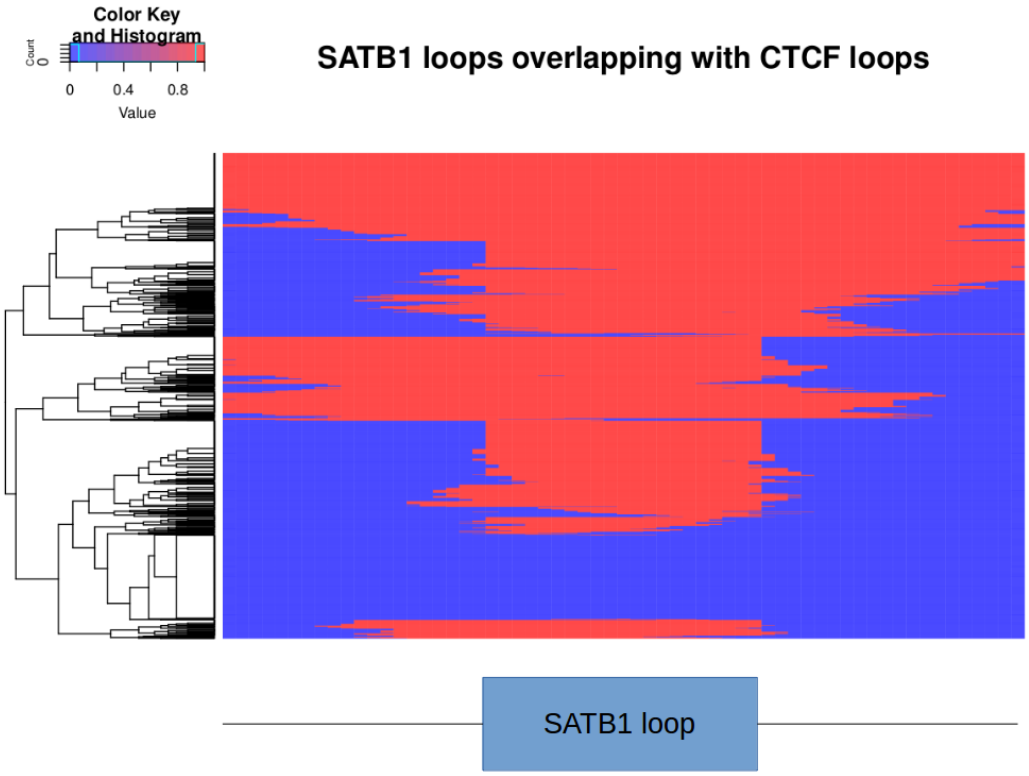
The two proteins though have different binding properties as shown by the nucleosome plots that we generated. SATB1 can probably bind nucleosomes (as Greenleaf suggested), while CTCF is found lying between nucleosomes:



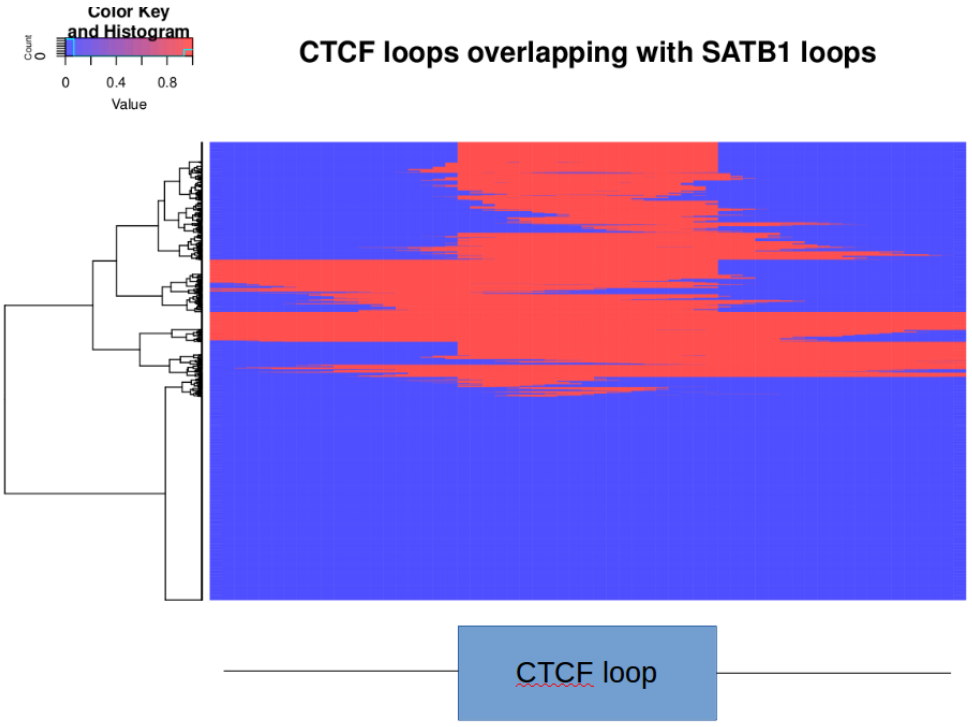
Nucleosome binding patterns across SATB1 and CTCF peaks. Nucleosomes were called using Nucleo-ATAC centered around peaks. Images created by Ilia Varamogianni.

SATB1 and CTCF associated loops are found in proximal areas of the genome

The loops isolated from the HiChIP experiment for each protein revealed that the loops associated with each factor lie extremely close : In a lot of cases you can find the same loop anchors for several called loops.



CTCF-SATB1 loops overlaps centered around SATB1 loops. Red depicts an overlap event: Thus for areas that are red throughout the SATB1 loop ‘’body’’ only, the corresponding SATB1 associated loop was similar with a CTCF associated loop.



It is evident that half of the CTCF loops are found to not be ‘’nested’’ in any way with SATB1 loops.

Regarding the loops :

CTCF loops (5k bin resolution) : **3029** loops , **1683** loops ‘’overlapped’’ with a SATB1 loop
SATB1 loops (5k bin resolution) : **1375** loops , **1139** loops ‘’overlapped’’ with a CTCF loop

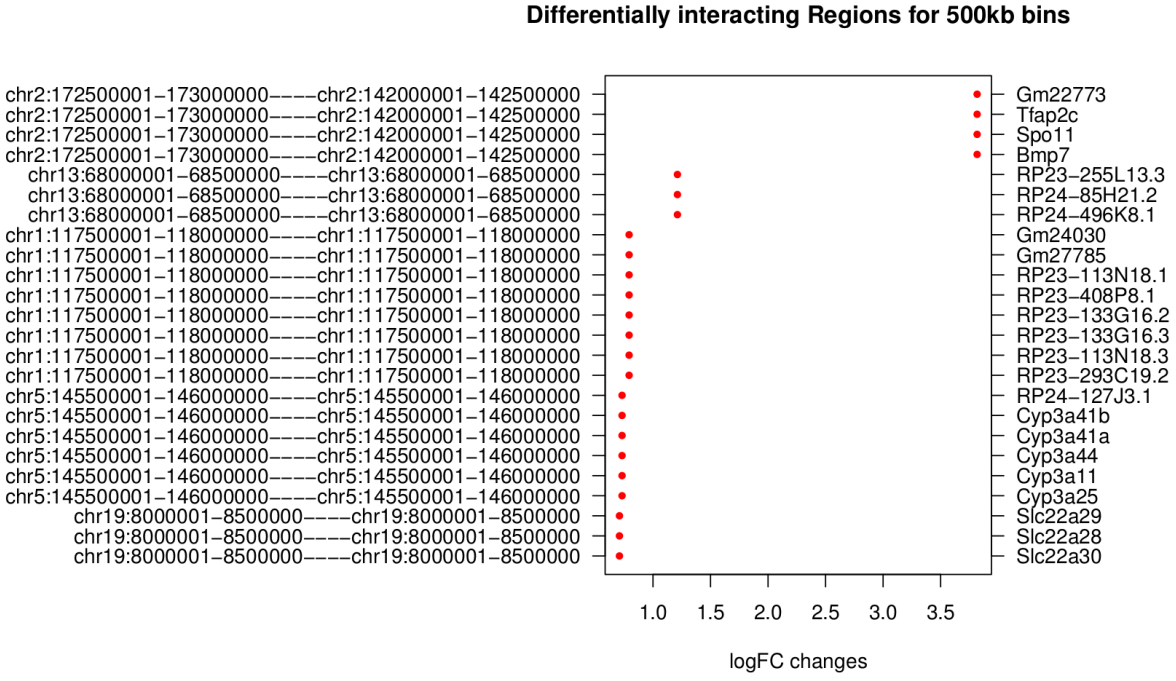
The ‘’overlapping’’ loops are enriched in immune-related genes, while the non-nested CTCF loops are enriched in more general categories (i.e. ras signaling).

CTCF- associated low resolution contact matrices exhibit stronger contact events than similar SATB1 matrices

It would be interesting to find out whether SATB1 could organize higher order 3D structures in a manner similar to CTCF. If such areas in the genome exist, then they would be enriched in SATB1 associated contact events. Since HiChIP for a specific protein enriches contact events associated with the corresponding protein, the contact matrices of this experiment can be used to answer the aforementioned question.

By comparing the contact matrices generated by the CTCF HiChIP and the SATB1 HiChIP experiments, differentially interacting areas can be isolated. SATB1 organized contact hubs can be isolated. DiffHic is an algorithm designed for such cases. However its main disadvantage is that it cannot be used for high resolution data. Thus diffHic was run for 100 kb and 500 kb resolution contact matrices.

The results seemed to suggest that at this resolution CTCF is the main genome organizer in thymocytes. **Only 7** interactions between bins were deemed as statistically significant (FDR <= 0.05) and were stronger in the SATB1 contact matrix.



Differential interactions isolated that are stronger in the SATB1 contact matrix. Only 7 interactions were isolated. The majority of the genes found in the above bins are not even expressed in thymocytes (e.g. Cypa family genes).

42 interaction pairs were deemed as statistically significant and were stronger in the CTCF contact matrix.

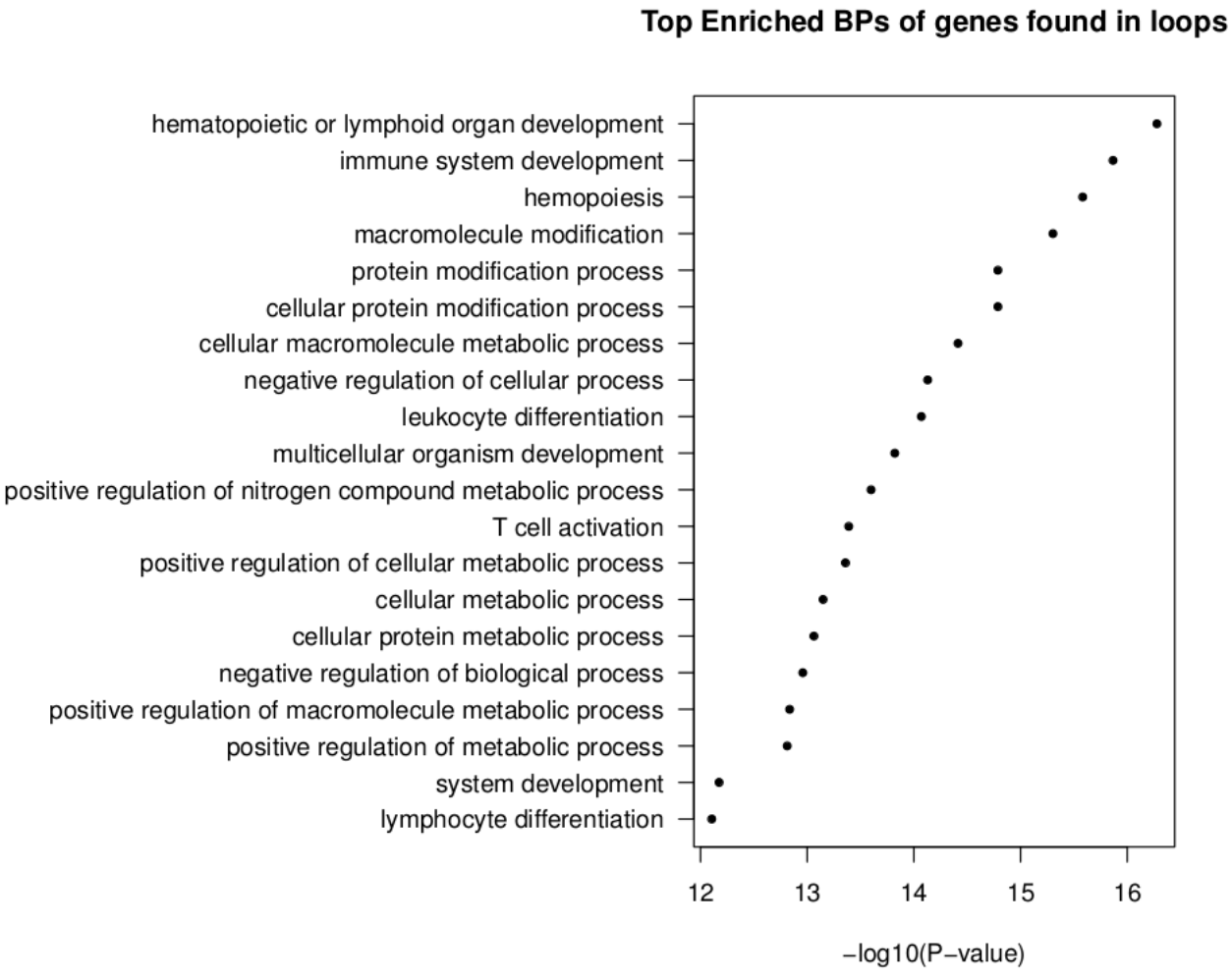
The results for the 100kb bin resolution were similar :

46 interaction pairs were found to be stronger in the SATB1 contact matrix.
553 interaction pairs were found to be stronger in the CTCF contact matrix.

It appears that CTCF is the main protein found associated with long-range contact events across the genome. Nevertheless these results describe low-resolution contact matrices. It would be interesting to find out whether SATB1 plays a role in forming loops between proximal sites in the genome. It appears though that higher order structures are mainly CT CF-associated.

Immune related genes are enriched in SATB1 – associated loops

Despite CTCF's dominance regarding looping and contact events, a significant over-representation of immune-related genes found in SATB1 associated loop-anchors can be observed.



The SATB1 anchors are also enriched for differentially expressed genes pointing out at possible regulation events :

Female, whole thymus RNA-seq

Underexpressed genes : 149/1436 found in anchors
Expected anchor – gene overlap : ~ 50 genes

Overexpressed genes : 76/1001 found in anchors
Expected anchor – gene overlap ~ 35 genes

Male, sorted Double Positive cells RNA-seq

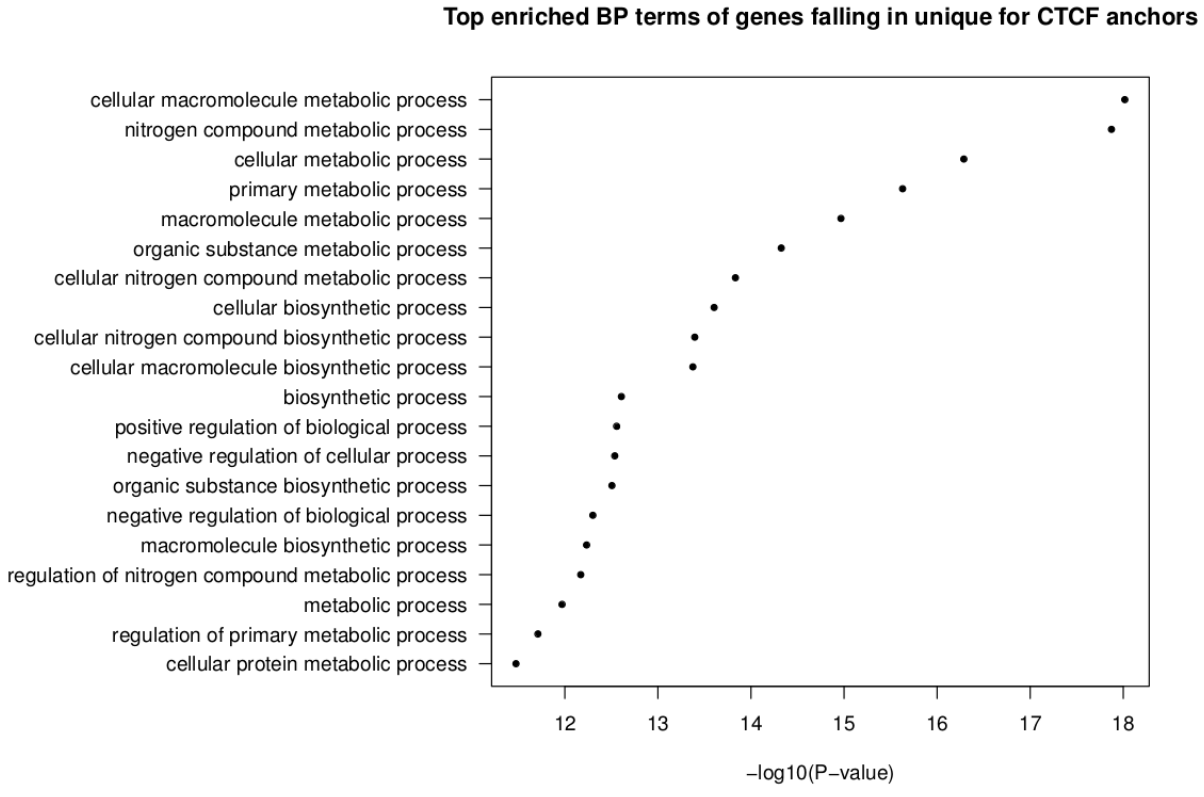
Underexpressed genes : 105/362 genes are found in anchors
Expected anchor – gene overlap ~ 15 genes

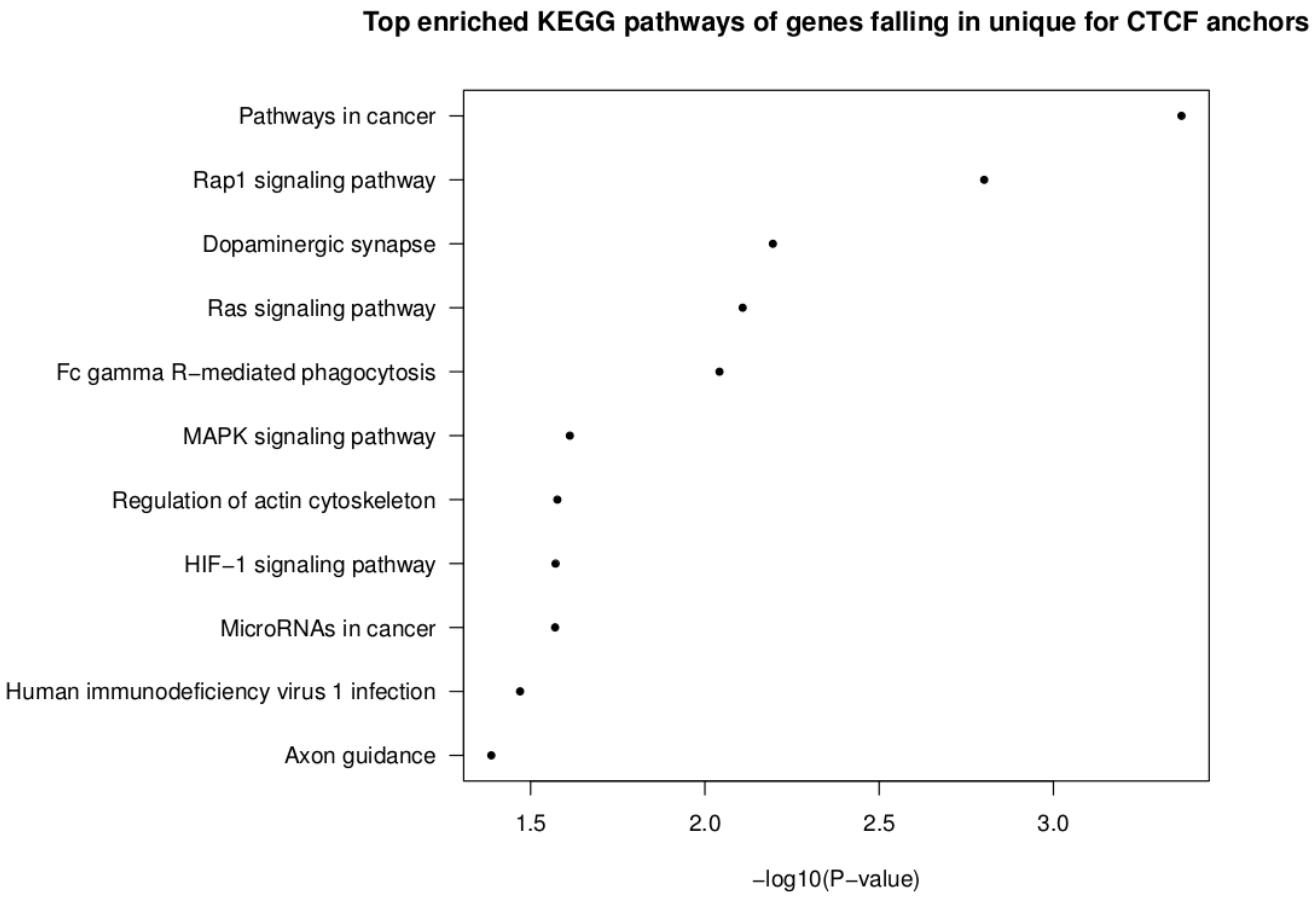
Overexpressed genes : 64/440 genes are found in anchors
Expected anchor – gene overlap ~ 19 genes

The above enrichments are higher for underexpressed genes, indicating that SATB1 loops mainly act as activators. **The same result is pointed out by the linear regression model that was constructed.**

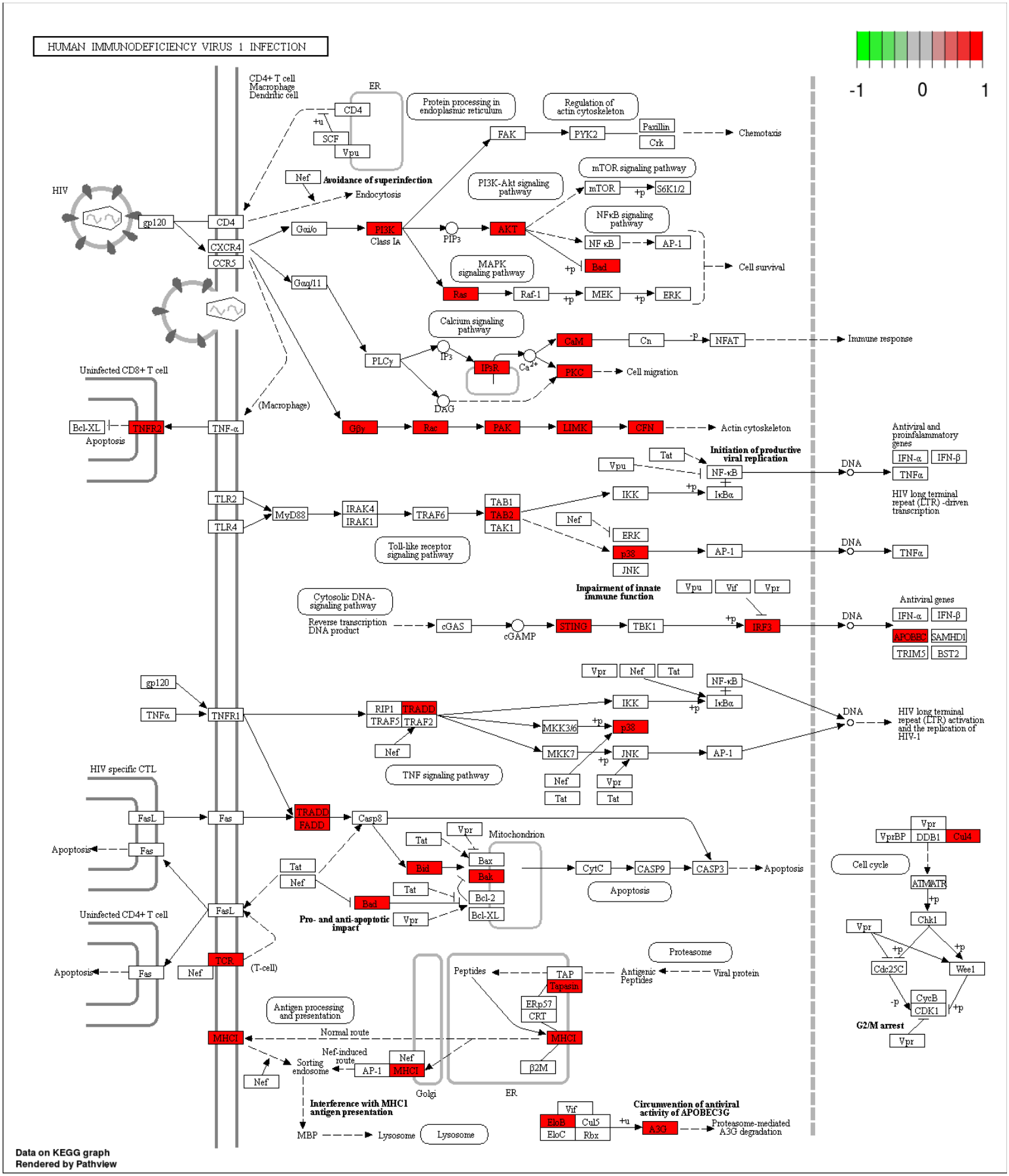
An open question remains : What are loops needed for? Given the dramatic phenotype of the *Satb1* cKO mice and the immune-specificity of genes inside loops, loops could in theory be affecting the expression of genes in various ways.

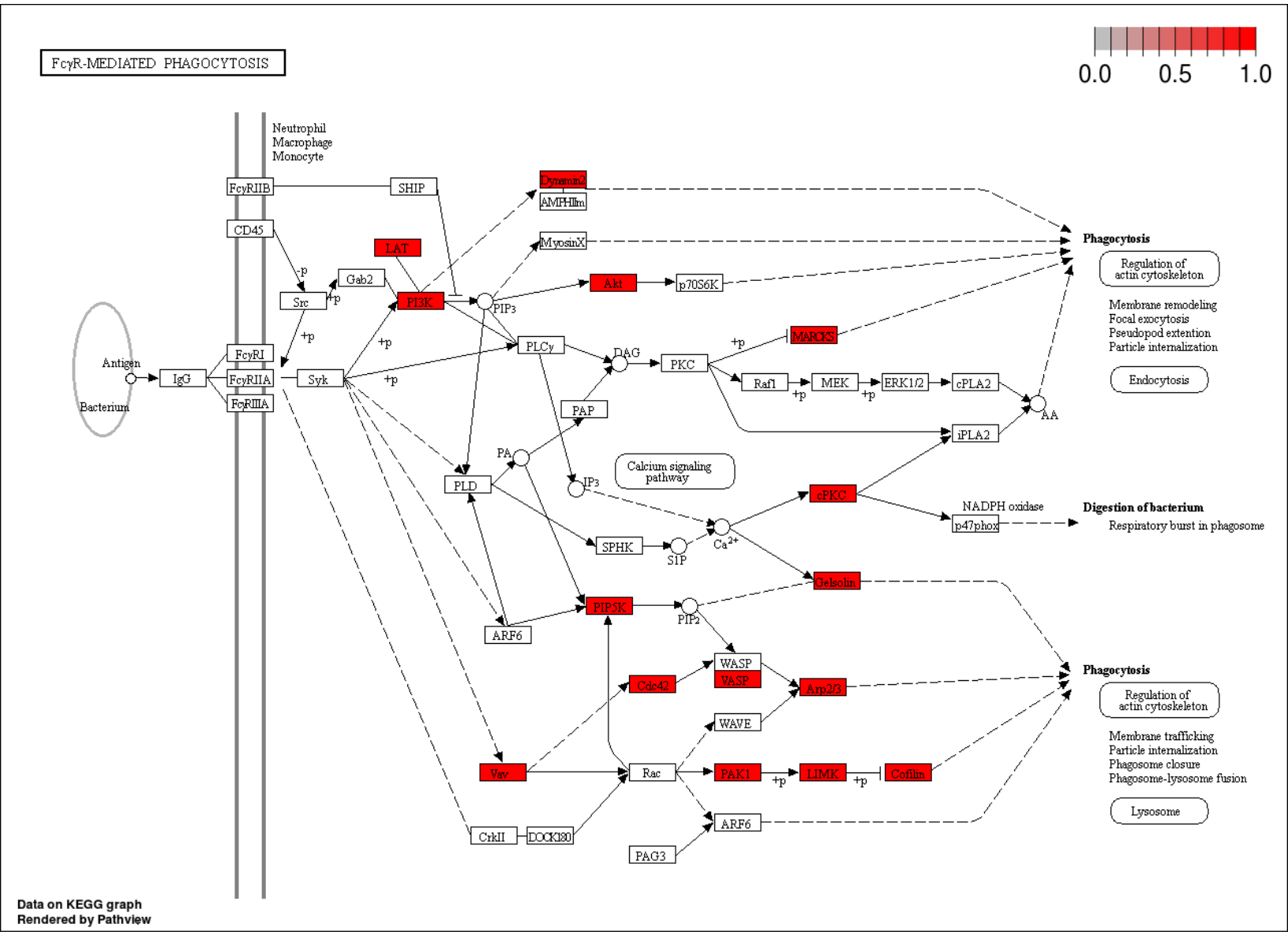
Finally functional analysis was performed on the isolated genes that resided in the anchors of CTCF associated loops that showed no overlap with the SATB1 associated loops. The genes residing in these loops showed enrichments for more general pathways.





Some immune KEGG pathways are also enriched for these genes, the “Human immunodeficiency virus infection” pathway and the “Fc gamma R-mediated phagocytosis”. The genes falling in each category were plotted on the corresponding pathway maps.


HUMAN IMMUNODEFICIENCY VIRUS 1 INFECTION
-1
0
1
CD4+ T cell
Macrophage
Dendritic cell
ER
CD4
+u
SCF
Vpu
Protein processing in endoplasmic reticulum
Regulation of actin cytoskeleton
FAK
PYK2
Paxillin
Crk
Chemotaxis
HIV
gp120
CD4
CXCR4
CCR5
Nef
Avoidance of superinfection
Endocytosis
Gαi/o
PI3K
Class IA
PIP3
AKT
PI3K-Akt signaling pathway
mTOR signaling pathway
mTOR
+p
S6K1/2
NFκB signaling pathway
NF κB
AP-1
Bad
MAPK signaling pathway
Cell survival
Gαq/11
Ras
Raf-1
MEK
ERK
Calcium signaling pathway
PLCγ
IP3
IP3R
Ca2+
CaM
Cn
-p
NFAT
Immune response
PKC
Cell migration
DAG
Uninfected CD8+ T cell
(Macrophage)
Bcl-XL
Apoptosis
TNFR2
TNF-α
Gβγ
Rac
PAK
LIMK
CFN
Actin cytoskeleton
Antiviral and proinflammatory genes
IFN-α
IFN-β
TNFα
Tat
Vpu
Initiation of productive viral replication
NF-κB
IκBα
DNA
HIV long terminal repeat (LTR) -driven transcription
TLR2
TLR4
MyD88
IRAK4
IRAK1
TRAF6
TAB1
TAB2
TAK1
IKK
Nef
ERK
p38
JNK
AP-1
Toll-like receptor signaling pathway
Impairment of innate immune function
Vpu
Vif
Vpr
Cytosolic DNA-signaling pathway
Reverse transcription DNA product
cGAS
cGAMP
STING
TBK1
IRF3
Antiviral genes
APOBEC
SAMHD1
TRIM5
BST2
gp120
TNFα
TNFR1
RIP1
TRADD
TRAF5
TRAF2
MKK3/6
MKK7
Vpr
Nef
Tat
HIV long terminal repeat (LTR) activation and the replication of HIV-1
TNF signaling pathway
HIV specific CTL
FasL
Fas
Apoptosis
TRADD
FADD
Casp8
Mitochondrion
Bax
Bak
Bid
Bad
Bcl-2
Bcl-XL
CytC
CASP9
CASP3
Apoptosis
Pro- and anti-apoptotic impact
Uninfected CD4+ T cell
TCR
(T-cell)
Proteasome
Peptides
TAP
Tapasin
Antigenic Peptides
Viral protein
ERp57
CRT
Antigen processing and presentation
MHCI
Normal route
Nef-induced route
Sorting endosome
AP-1
Golgi
ER
β2M
Interference with MHC1 antigen presentation
MBP
Lysosome
Vif
EloB
EloC
Cul5
Rbx
A3G
Circumvention of antiviral activity of APOBEC3G
Proteasome-mediated A3G degradation
VprBP
DDB1
Cul4
Cell cycle
ATM&ATR
Chk1
Cdc25C
Wee1
CycB
CDK1
G2/M arrest
Data on KEGG graph
Rendered by Pathview

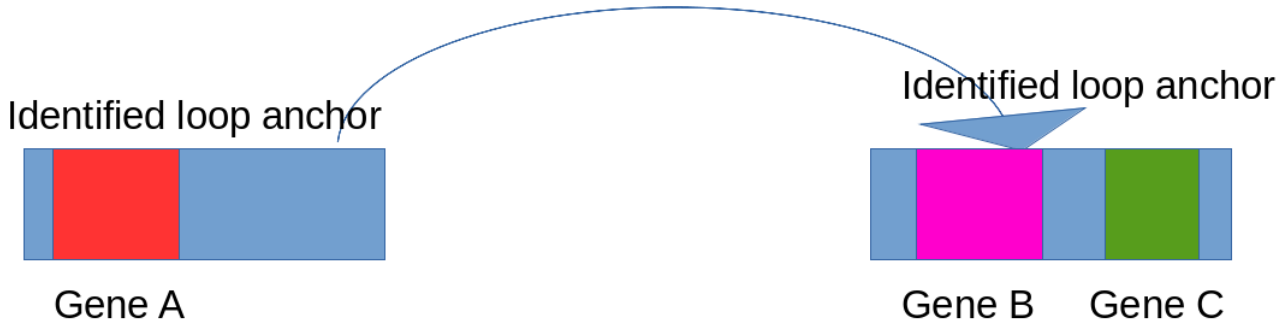


Both of the enriched pathways include various component of the *Akt* pathway. The "FcγR – Mediated Phagocytosis" pathway is also containing a lot of genes regulating the actin cytoskeleton. Thus this pathway seems to not be directly related to thymocytes.

The "HIV1 virus infection" KEGG pathway contains a lot of *Akt* pathway genes, as well as cytoskeleton-related genes. However it seems that a small subset of genes are immune related : MHC-related genes are found in unique for CTCF anchors, along with *Tradd* and *Fadd* which are members of the *Tnf* signaling pathway. Although the *Tnf* signaling pathway is a crucial player for immune responses, it is an active pathway that is relevant for multiple different cell lineages. All in all it seems that CTCF associated loops, that are devoid of SATB1 loops, are found in genes that are related to ubiquitous pathways.

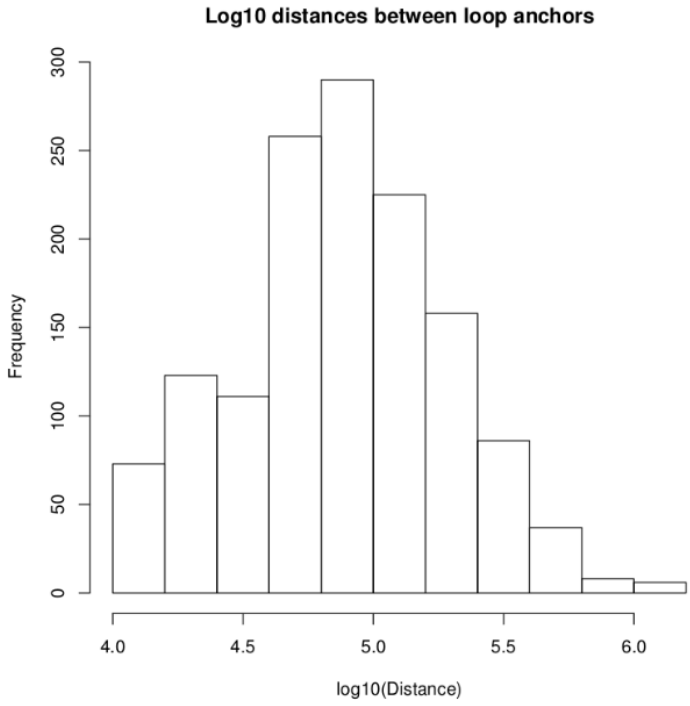
Genes inside SATB1 loops and SATB1 anchors do not share similar expression levels

Loops, or loop anchors, could function as transcriptional hubs, containing genes that are expressed in similar levels. To test this hypothesis the following approach was employed :

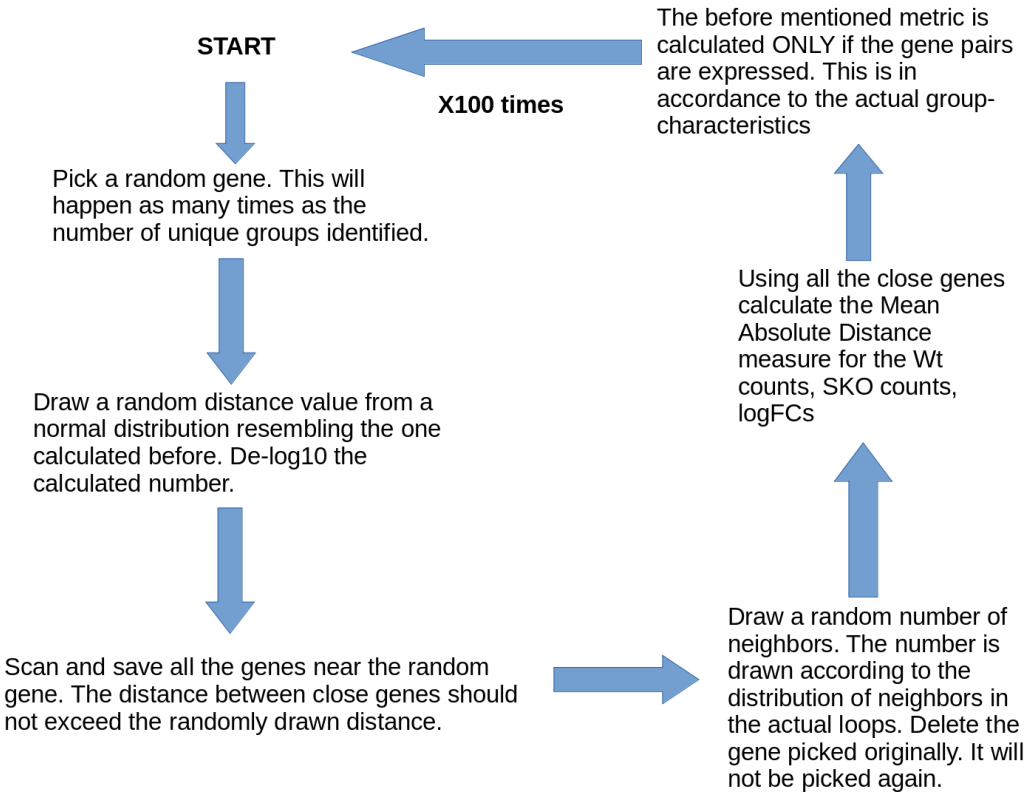


We isolated all gene pairs that resided in loop anchors and are connected via SATB1 loops. We calculated the mean Manhattan distance between their expression values.

We calculated the above distance measures in all **unique** gene groups identified : If the same group of genes is connected via different loops, it will only be considered once. Moreover the expression patterns of genes in groups were isolated. It seems that very few genes connected via loops (only 10) are not expressed. The distribution of distances between anchors is the following one :



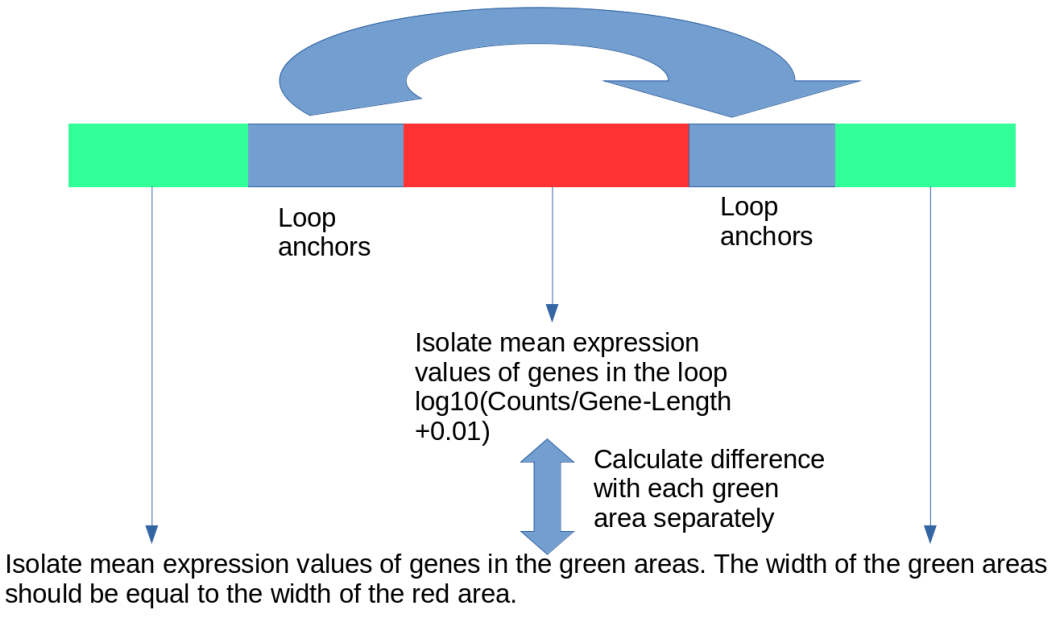
In order to compare the above distribution of values, a complex null distribution was constructed using the following procedure :



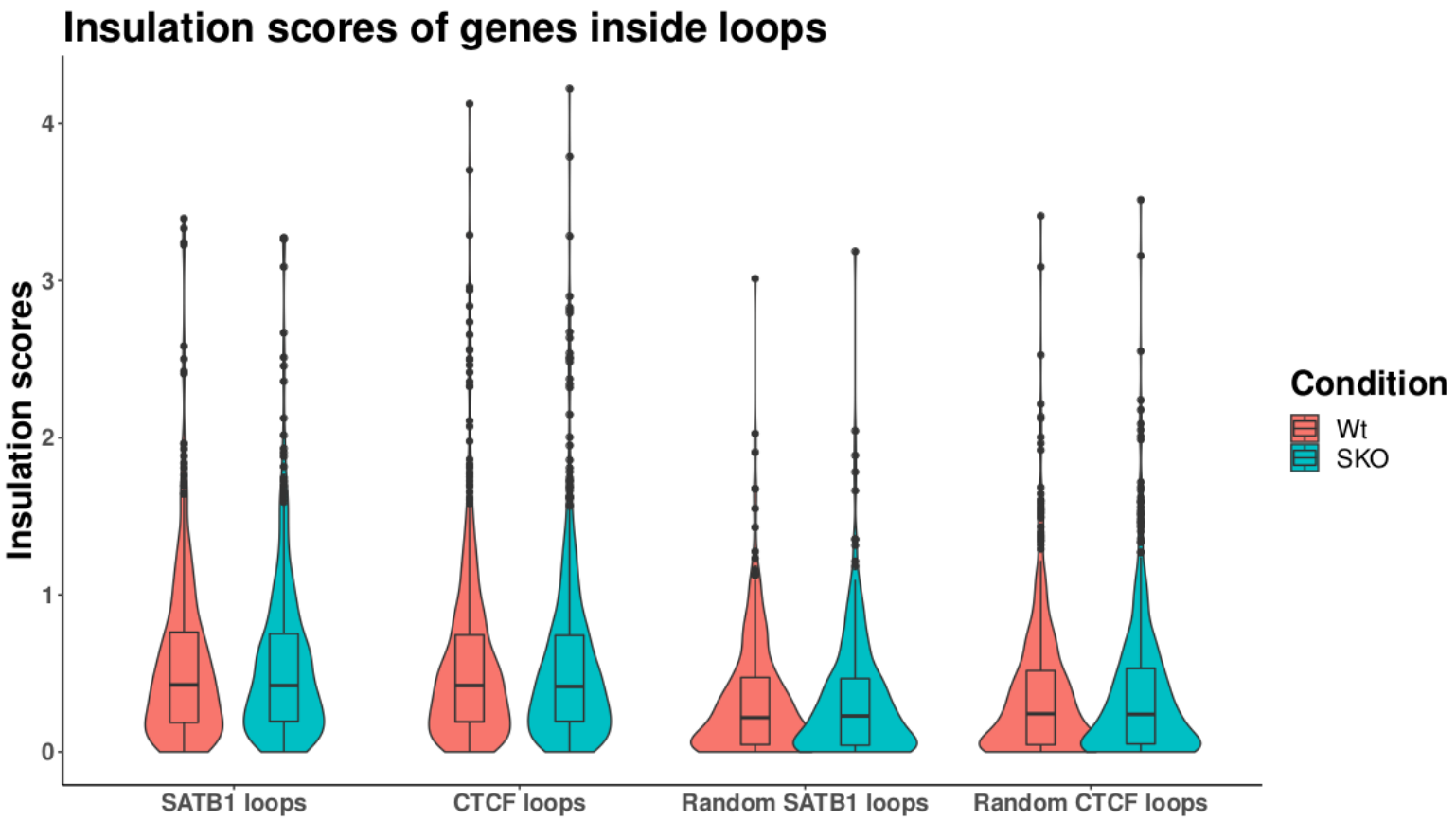
The results indicated that genes in anchors **do not** exhibit similar expression levels. The expression distance metric employed was similar between the null distribution and the different samples (Wt expression values, *Satb1* cKO expression values). This was also proven to be true for genes that resided inside SATB1 loops.

<u>The expression values of genes inside loops differ with those of their neighbors</u>

Another possible functional use of looping would be to isolate the specific genes from their neighborhood. In that case, the expression levels of genes that resided inside loops should differ with those that are outside the loop but in close proximity. To test this hypothesis, the following steps were performed :



Using the above procedure, we isolated the expression differences of genes inside loops in regard to their neighbors. To compare the above metric, randomly isolated gene neighbors were drawn at random, and the above scores were calculated for each randomly generated neighbor. The results are depicted below :



It seems that genes inside loops have bigger differences in their expression levels in regard to their neighbors. The discrepancy here is that this phenomenon appears to be independent of the presence of SATB1. The same genes in the *Satb1* cKO seem to preserve this tendency. The most obvious explanation is the presence of CTCF near the SATB1 associated loops.

SATB1 associated loops are enriched for enhancer-promoter connections in immune-related genes

A final role for loops could be the establishment of connections between promoters and regulatory regions. Enhancers serve as a paradigm for such connections, being capable of enhancing transcription of genes under their control. Moreover, enhancers share characteristic epigenetic marks making their mapping across specific tissues feasible (in contrast to silencers for example).

The Bing Ren lab, has mapped enhancer coordinates in the thymus, using a hidden-Markov model. Enhancers were evaluated in the basis of having the H3k4me1 epigenetic mark but not the H3k4me3 epigenetic mark, which reflects active transcription. Using known enhancer locations in embryonic stem cells in addition to the epigenetic marks described, the hidden-Markov model was trained in this

context and was later used in order to isolate thymus-specific enhancers. Only the center of the **5605** enhancers was reported. Thus the enhancers were extended 50 bp upstream and downstream simultaneously.

If SATB1 serves as a connector between enhancers and target genes, then SATB1 should bind the enhancers sequences. Using the Chip-seq experiment for SATB1 the following overlaps were calculated :

SATB1 Chip-seq (Krangel's dataset) → **913** overlap events
Expected overlap based on a permutation analysis → 76 overlap events (~11 fold enrichment)

SATB1 Hi-ChIP peaks → **68** overlap events
Expected overlap based on a permutation analysis → 14 overlap events (~ 5 fold enrichment)

CTCF Chip-seq → **427** overlap events
Expected overlap based on a permutation analysis → 100 overlap events (~4.25 fold enrichment)

In general all the organizers seem to bind a lot of enhancer coordinates. It remains to be seen whether enhancers are also enriched within loop anchors. The respective calculations were once again performed :

SATB1 loop anchors (5k bin resolution) → **208** unique anchors overlapped with an enhancer
Expected overlap based on a permutation analysis → 22 overlaps (~ 11 fold enrichment)

CTCF loop anchors (5k bin resolution) → **264** unique anchors overlapped with an enhancer
Expected overlap based on a permutation analysis → 65 overlaps (~4 fold enrichment)

It seems that SATB1 loop anchors ''contain'' a lot of enhancers. The SATB1 enrichment is much higher than the CTCF one. It is known that CTCF does not tend to bring enhancers close to promoters : Its enrichments near enhancer sequences is due to its spatial proximity to formed loops (Schoenfelder et al. ,2019).

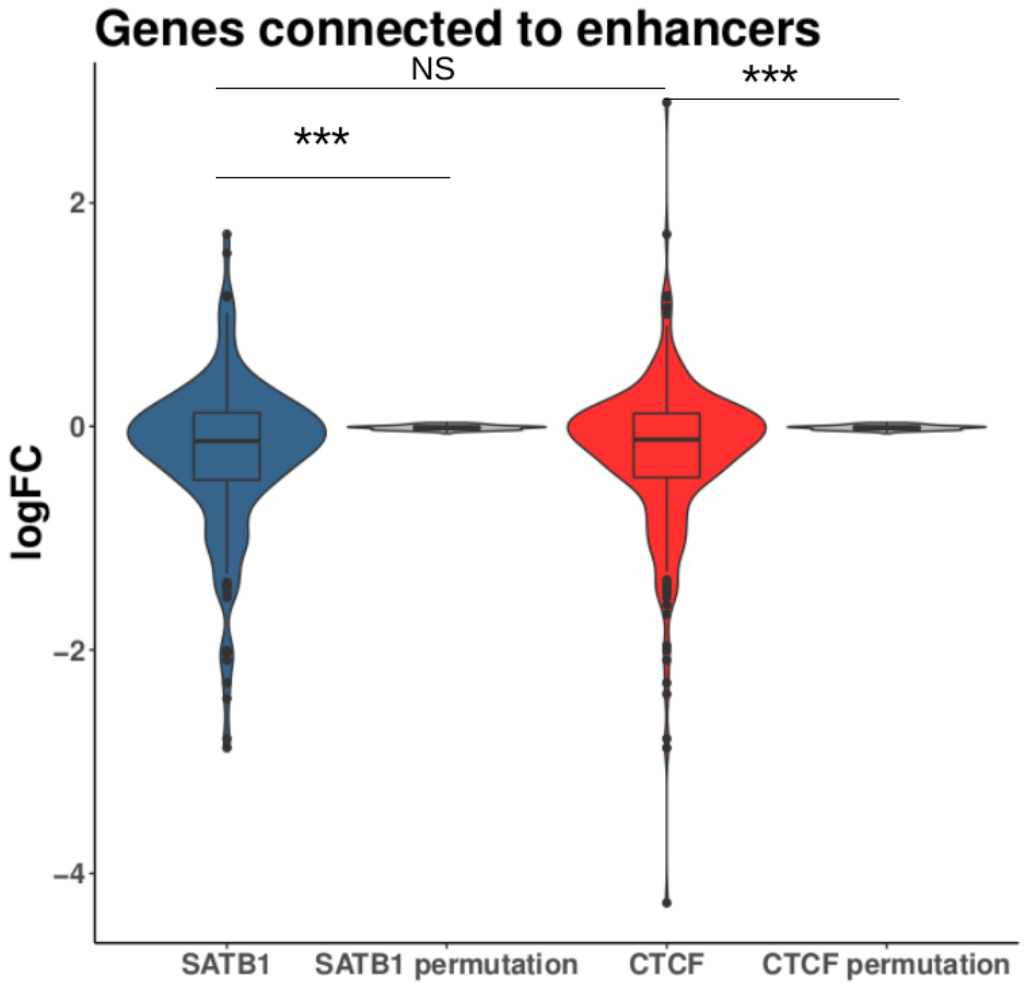
Genes that were connected to an enhancer via a SATB1 associated loop were isolated. A total of 192 genes were isolated. By shuffling the loop coordinates (kept the distance between shuffled anchors to 100kb as the mean distance of the real SATB1 associated loops), only 8 genes showed this behavior. Genes that were connected to an enhancer via a CTCF associated loop were also isolated and were in total 302. A similar permutation analysis showed that 43 genes would be isolated if the CTCF loops were shuffled at random. This striking difference in enrichment values, highlights a potential SATB1-specific action regarding such connections.

If SATB1 is causal for the formation of loops between enhancer-promoters then we should expect the following tendencies :

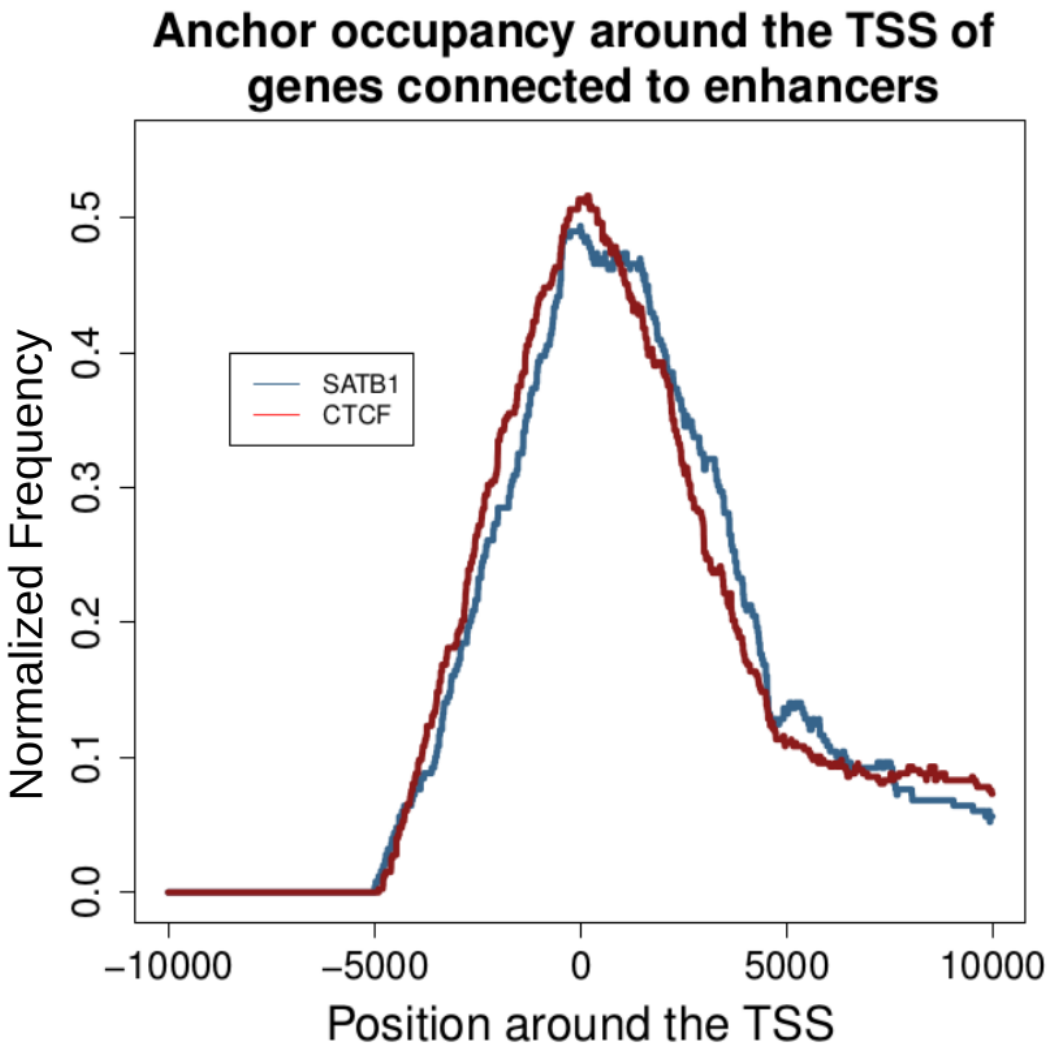
a) In the absence of SATB1 the genes that were connected to an enhancer via a SATB1 loop should have lower expression levels.

b) The enhancer-promoter connection should occur near the transcription start site of the gene.

It truly seems that in the absence of SATB1, there is a negative trend in expression values. On average, genes that were connected to an enhancer via a SATB1-associated loop drop their expression values by 16% (mean logFC is -0.2475). A permutation analysis showed that this drop is statistically significant and negatively enriched :

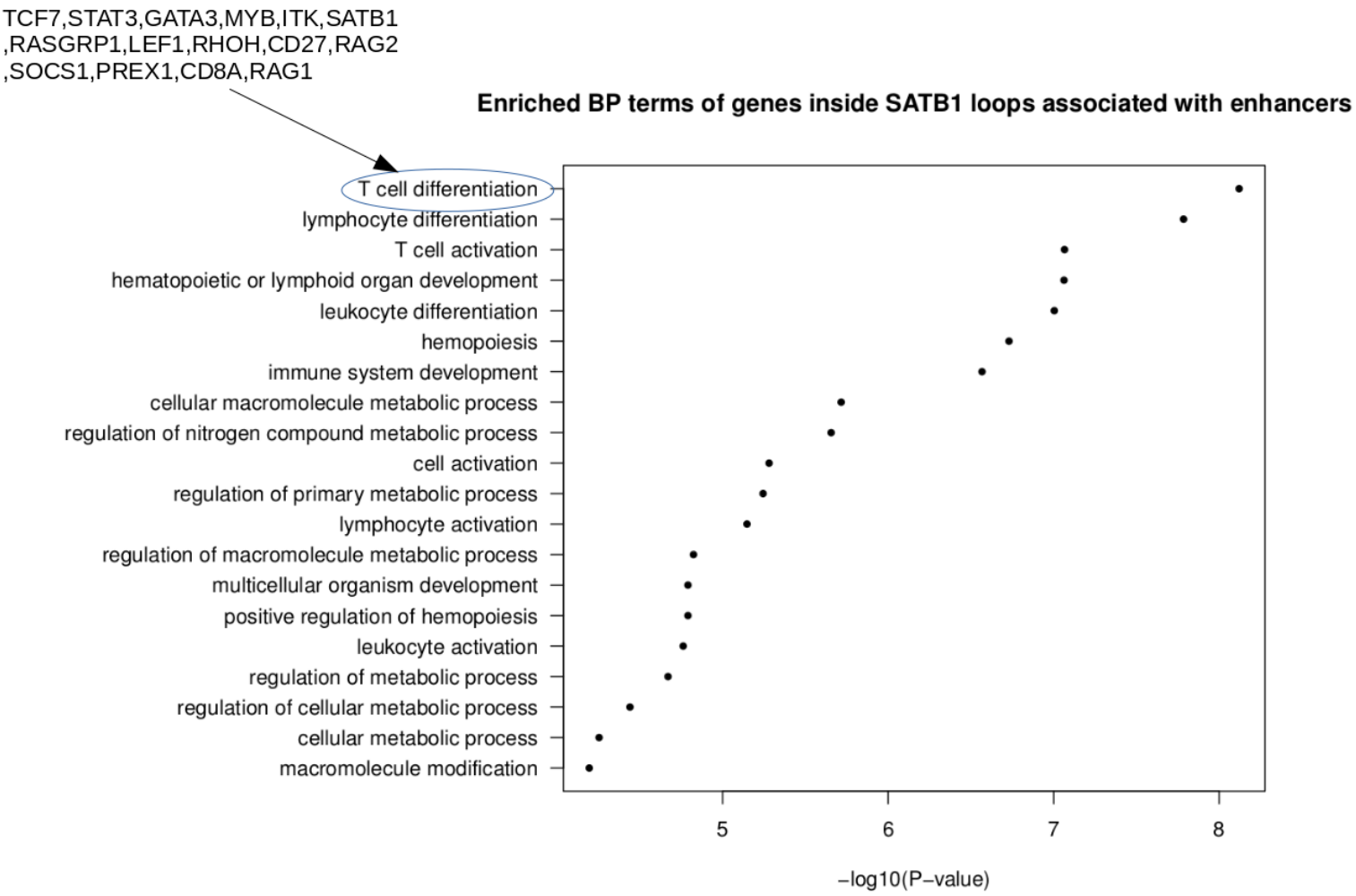


Moreover average gene plots were constructed for the anchor occupancy along the transcription start sites of genes that were connected to enhancers. If an enhancer-promoter connection is established, then we should expect that the loop anchor occupying the gene, should be proximal to its transcription start site. This appears to be the case :



Worth noting is that the anchors are 5kb bins. It seems that the majority of the anchors for the isolated genes reside in areas proximal to the TSS.

An open question remains about the nature of the genes found in such enhancer-promoter connections via SATB1 – associated loops. Functional analysis of the genes revealed an enrichment in immune related pathways as shown below. An interesting finding is that the genomic locus of *Satb1* itself, was found to be connected to an enhancer via a SATB1-associated loop. This could point out to a self-regulatory positive feedback loop.

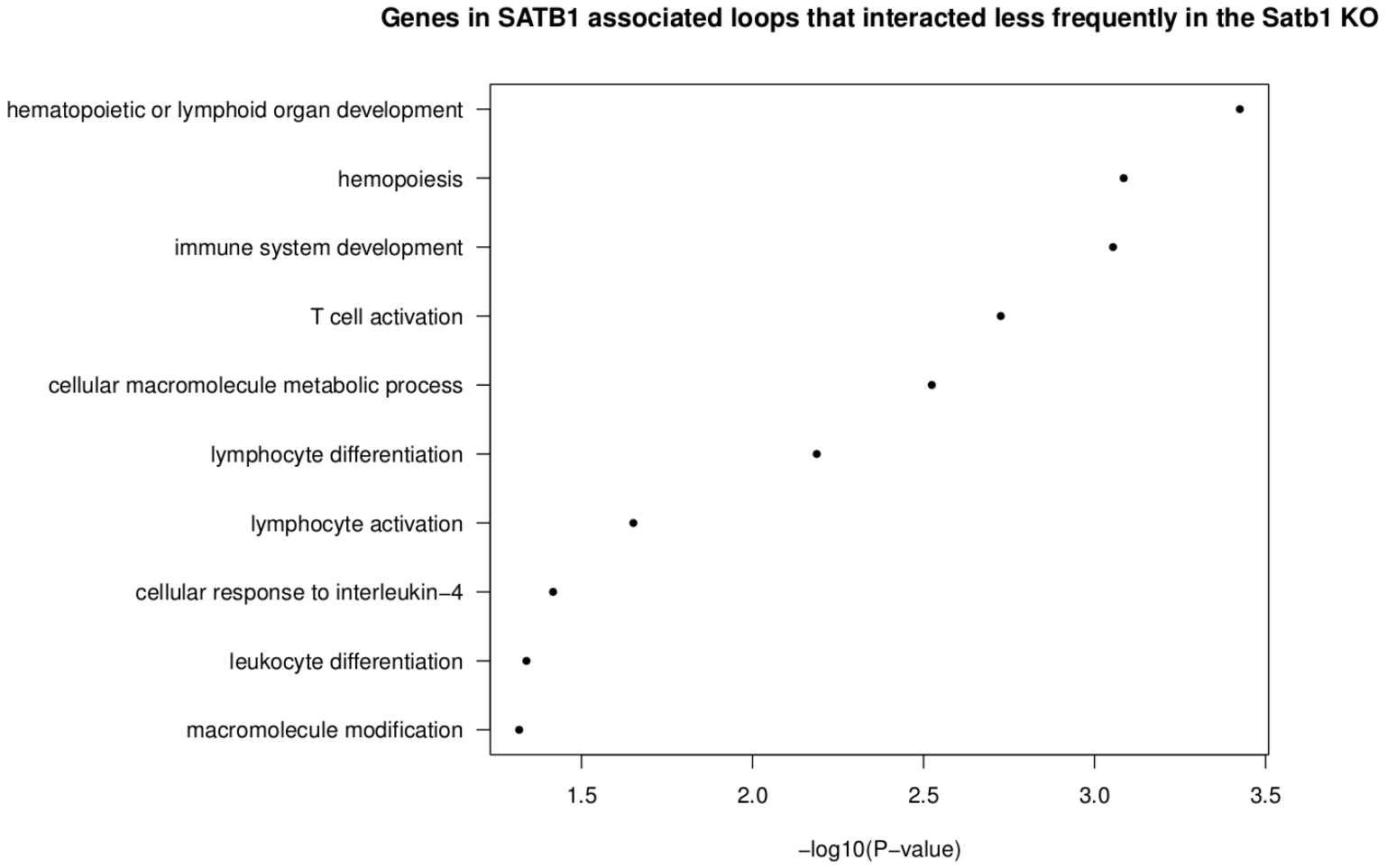


The genes that were found in the first enriched BP-term are crucial for proper T-cell development. The *Rag1* and *Rag2* genes are known targets of SATB1's regulation via the creation of a loop between their promoter with a cis-regulatory element. *Tcf7, Lef1 , Cd8a* are also genes implicated in T-cell development and most importantly lineage decision (CD4 single positive cells versus CD8 single positive cells).

A subset of SATB1 associated loops are reported as less interacting H3k27ac loops in the *Satb1* cKO

The presence of SATB1 in a specific loop doesn't necessary translate to a loop being formed by SATB1. A lot of examples with proteins ''hijacking'' created loops are well documented (e.g. p53) (Schoenfelder et al. ,2019). In order to delineate whether SATB1 is involved in the establishment of these loops HiChIP experiments for the epigenetic mark H3k27ac were conducted. H3k27ac is a mark found in active enhancers and in general it correlates with activate transcription. Thus if a loop is SATB1 mediated, then in the absence of SATB1 there should be less interactions between the corresponding loop anchors. FitHiChIP enables the identification of loops that are more or less interacting between two conditions. Thus H3k27ac loops that interacted **less in the *Satb1 cKO* were isolated.**

Out of 11540 isolated less interacting H3k27ac loops, 132 loops shared **exactly** the same coordinates with SATB1 – associated loops (in total 1375 SATB1-associated loops). 186 genes were found inside these common loops.



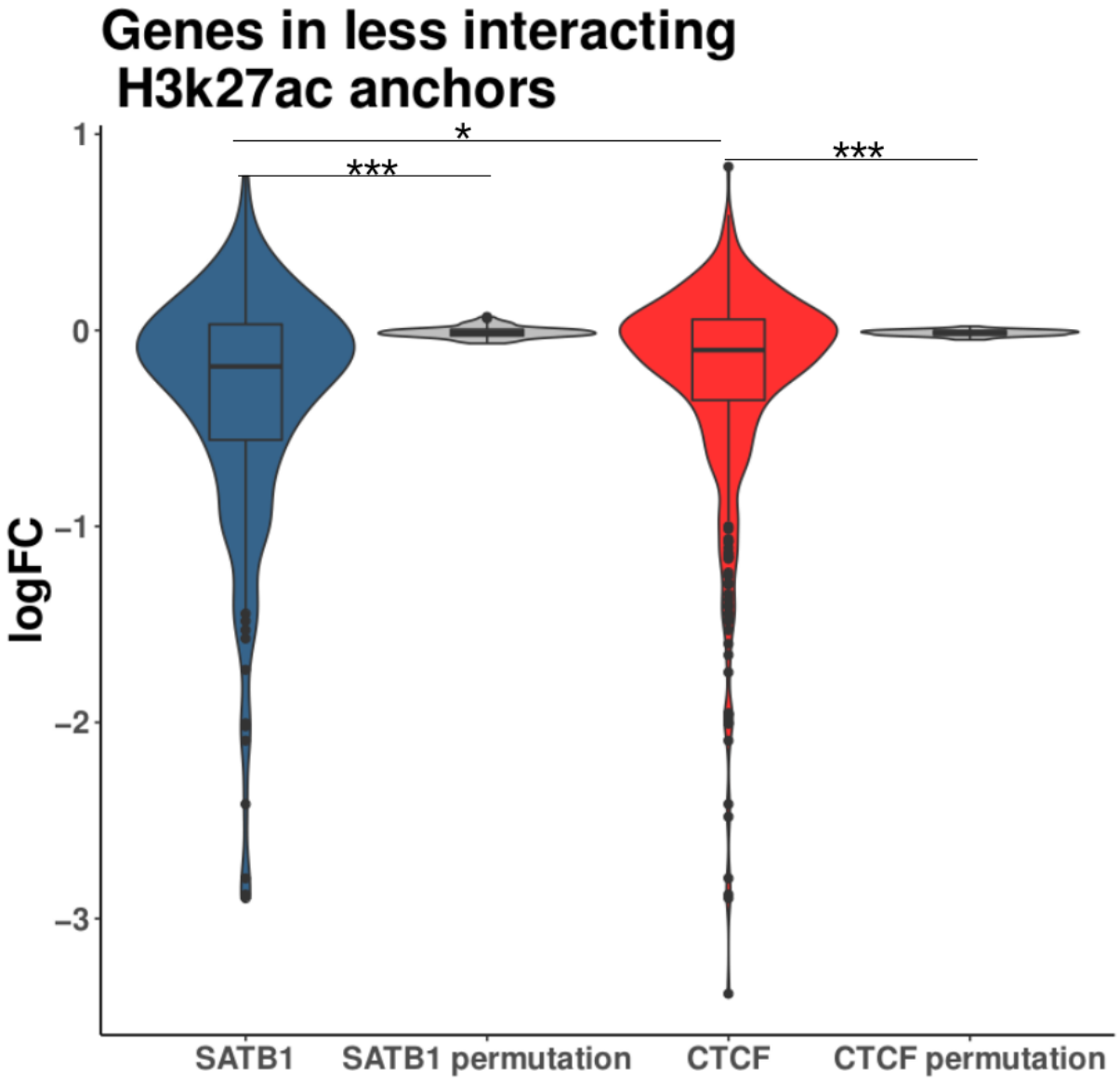
Once again, immune pathways - enrichments are evident.

The genes falling in the top BP term are :

Tcf7,Fli1,Myb,Mknk2,Itk,Runx1,**Lef1**,Wasf2,Lfng,Il4ra,Ly6d,Wdr78,Rps19,Pik3r1,Zfp36l2,Cd24a,**Cd8a**,Mapk3,Chd2

Some highlighted genes were presented in the previous enhancer analysis.

The isolated genes exhibit a significant average drop in expression levels (-22% or logFC -0.354) as is shown below :



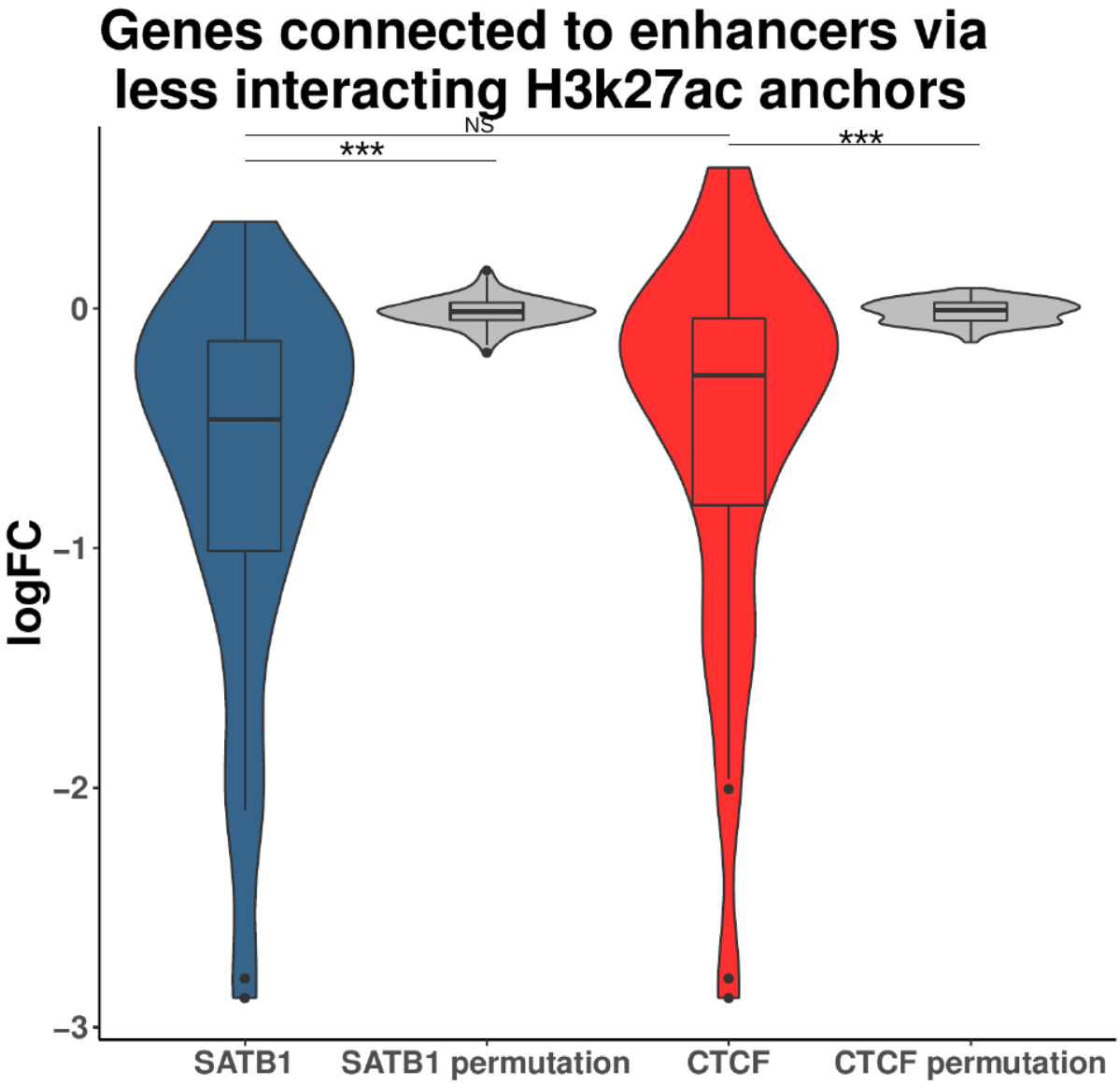
<u>Critical T-cell fate regulators are regulated via SATB1 mediated enhancer-promoter contacts</u>

Out of the above 132 loops isolated, only a subset connects genes with enhancers. 25 SATB1 associated loops that interacted less in the *Satb1* cKO were connecting genes with enhancers. Those loops mediated such enhancer contacts for 37 total genes. The isolated genes are the following :

Als2cl, Atl2, ***Cd8a,*** *Chrna9, Dusp4, Egln1, Ets2, Gtf2ird2, Itk, Kpna4,* ***Lef1,*** *Mgat1, Mier1, N4bp2 Rmnd5a, RP23-135L20.1, RP23-223A11.6, RP23-265F20.14, RP23-265F20.7, RP23-268N22.4 RP23-273O7.3, RP23-322E20.9, RP23-349J3.1, RP23-393I12.2, RP23-480M17.3, RP24-176P19.3 RP24-282D16.2, RP24-282D16.3, RP24-359H20.3, RP24-443G20.1, RP24-443G20.2, Runx1 Spsb1,* ***Tcf7,*** *Tha1, Wdr78, Zfp280d*

Once again *Tcf7*,*Lef1*, *Cd8a* appear to be regulated via SATB1. The genes are critical for the development of T-cells and SATB1 could directly regulate their expression level by forming loops between their promoter sites and enhancer sequences. The expression levels of the three above genes is lowered in the absence of SATB1, **in both RNA-seq datasets.**

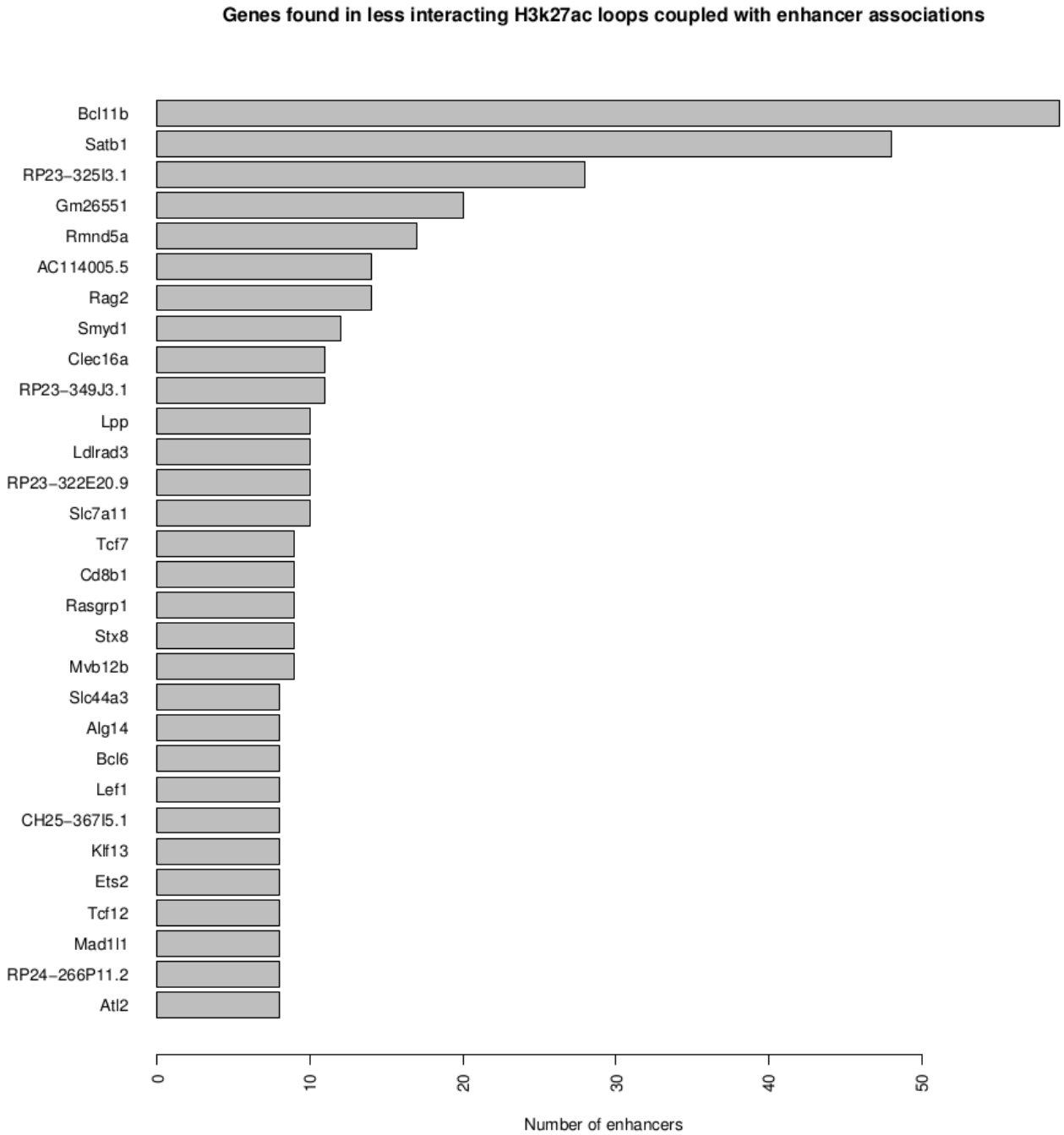
Regarding the expression levels of the aforementioned genes in general, there is a notable decrease in the absence of SATB1 :



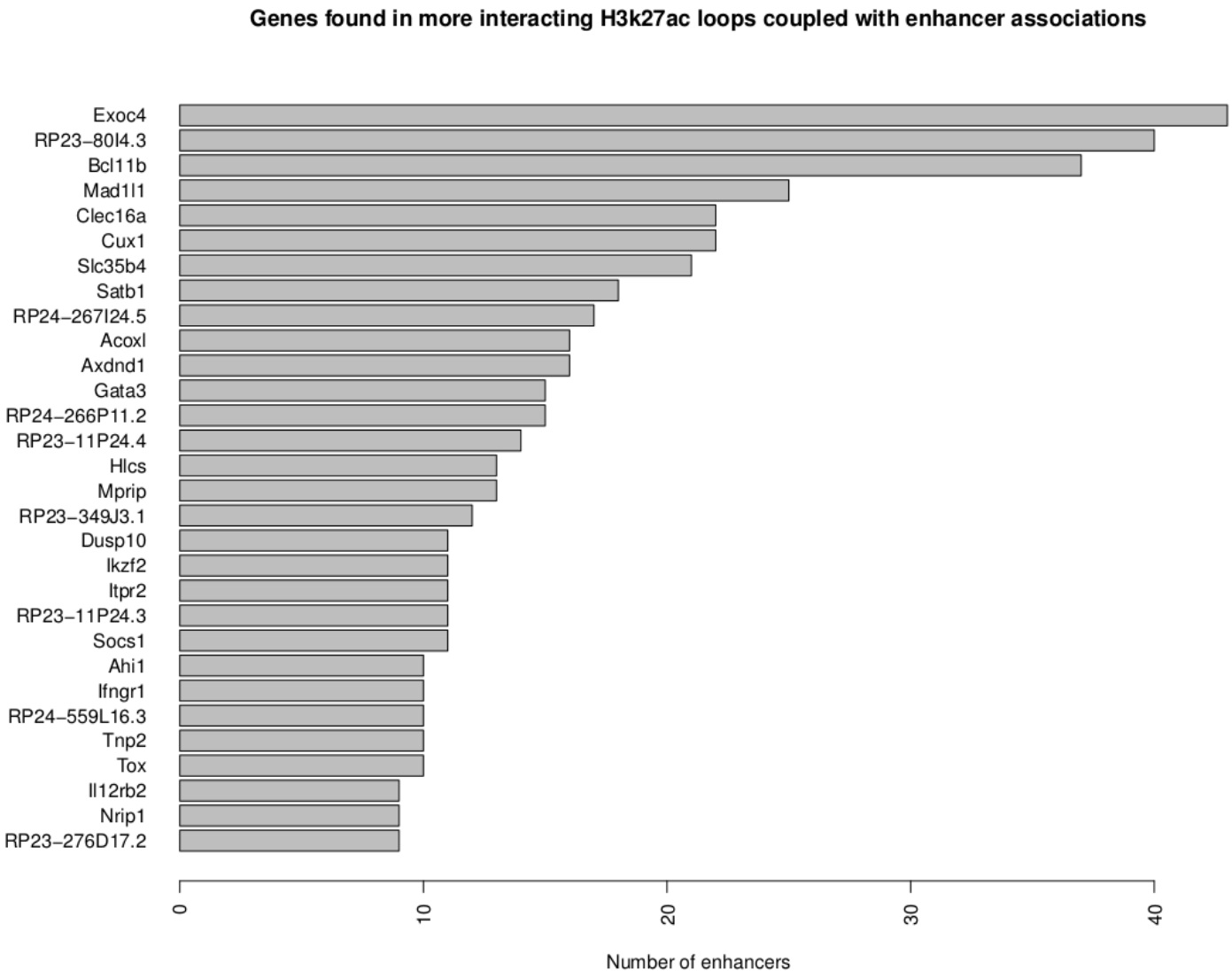
Enhancer – gene connections via H3k27ac loops correlate well with expression changes

In order to capture the general loop ‘’rewiring’’ associated with SATB1 ablation, a separate kind of analysis was carried out using the differentially interacting H3k27ac loops.

11540 H3k27ac loops were deemed as less interacting loops in the absence of SATB1, while 12111 loops were deemed as more interacting loops in the absence of SATB1. Enhancer-gene connections were once again isolated using the previously described enhancer dataset. The number of established enhancer – gene connections in the case of more interacting H3k27ac loops was isolated for each gene and was stored as a count file. The same procedure was followed for the lost enhancer-gene connections. The corresponding plots for the genes with the larger enhancer ‘’rewiring’’ are shown below :



The *Satb1* genetic locus once again appears to ‘’lose’’ a lot of enhancer associations, indicative of a possible self-regulated positive feedback loop. *Bcl11b,* a crucial transcription factor shows the same tendency as the *Satb1* locus.

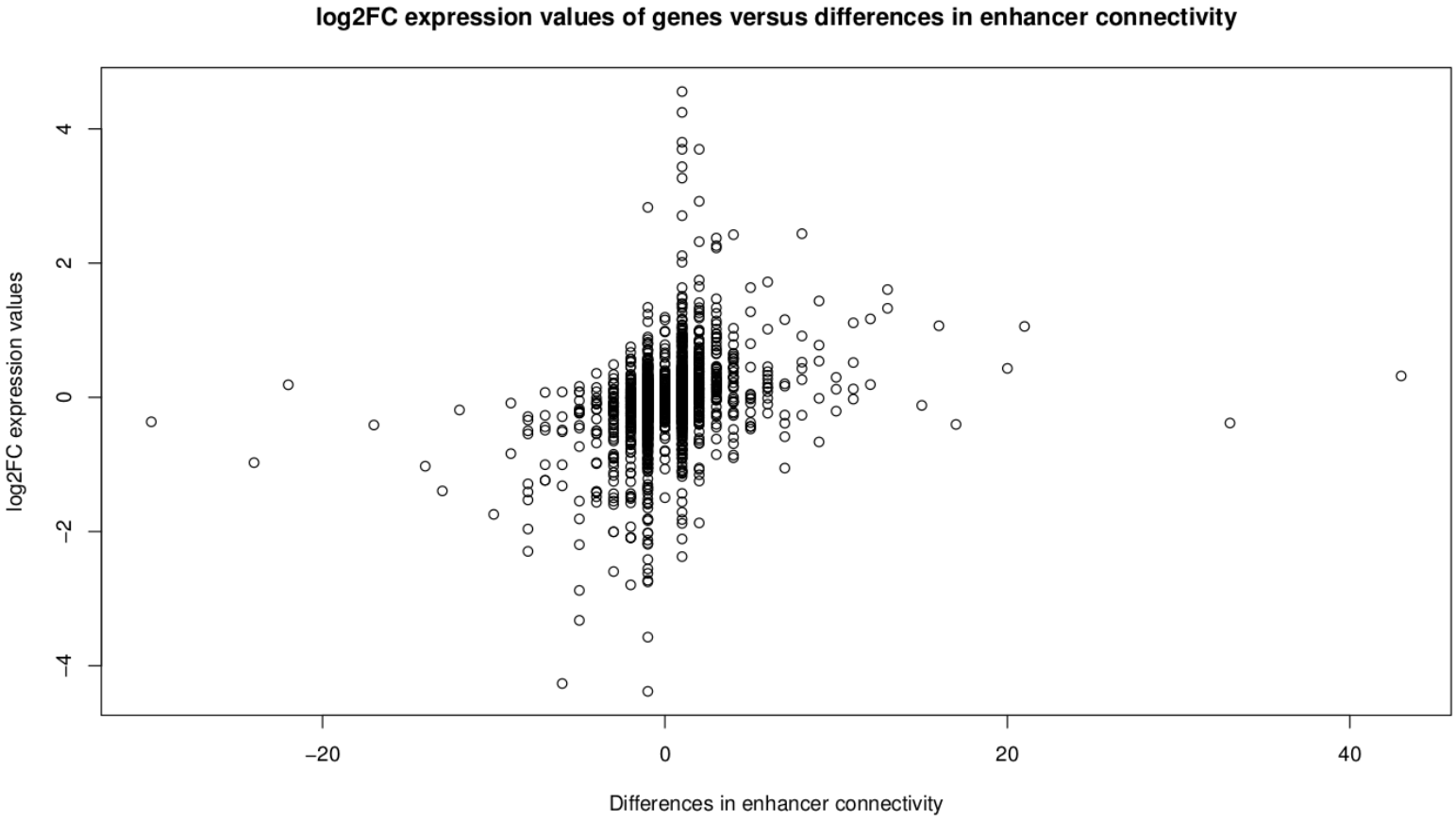


A discrepancy is evident. Common genes appear in the above plots. Bcl11b is a striking example and may be indicative of the complex regulation associated with specific genes.

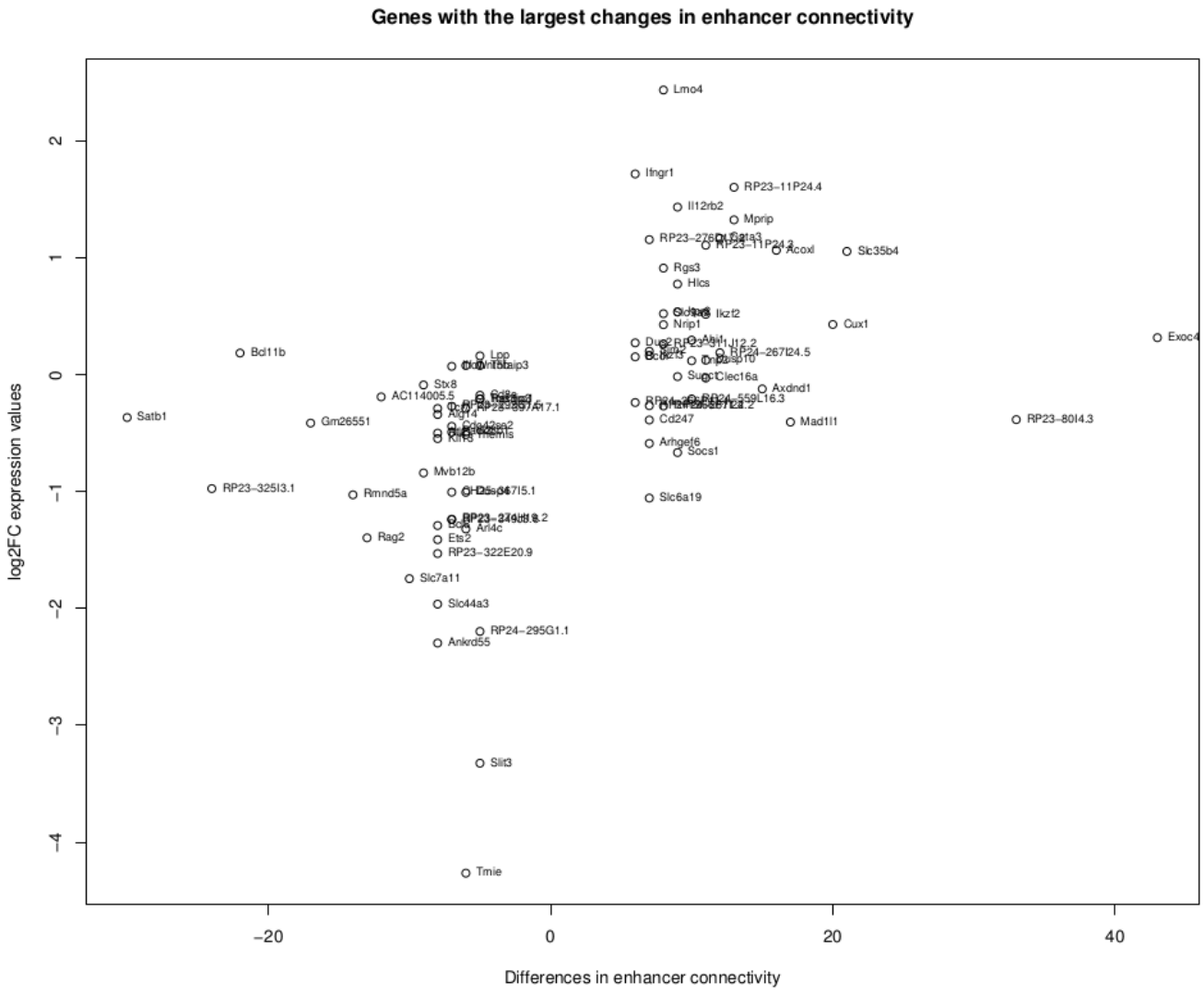
In order to study how gene expression is associated with enhancer gains or loses, a simple score was calculated for each gene i :

Enhancer net gain (i) = Enhancers gained (i) – Enhancers lost (i)

After the calculation of the above score for each gene, the logFC changes for each gene were also isolated. A scatter plot depicting the ‘’Enhancer net gain‘’ metric in association with the logFC change is show below :



A 0.34 **Spearman** correlation was calculated for the above values. A linear relationship is also evident in the above plot, even when using a simple score showing the differences in enhancer connectivity. The above plot was constructed again using only the genes exhibiting the largest differences in enhancer connectivity.



The Spearman correlation is 0.63 for the above genes. Several interesting genes are depicted like the *Ifngr1* which codes for the interferon-receptor and could be further studied by biochemical means.

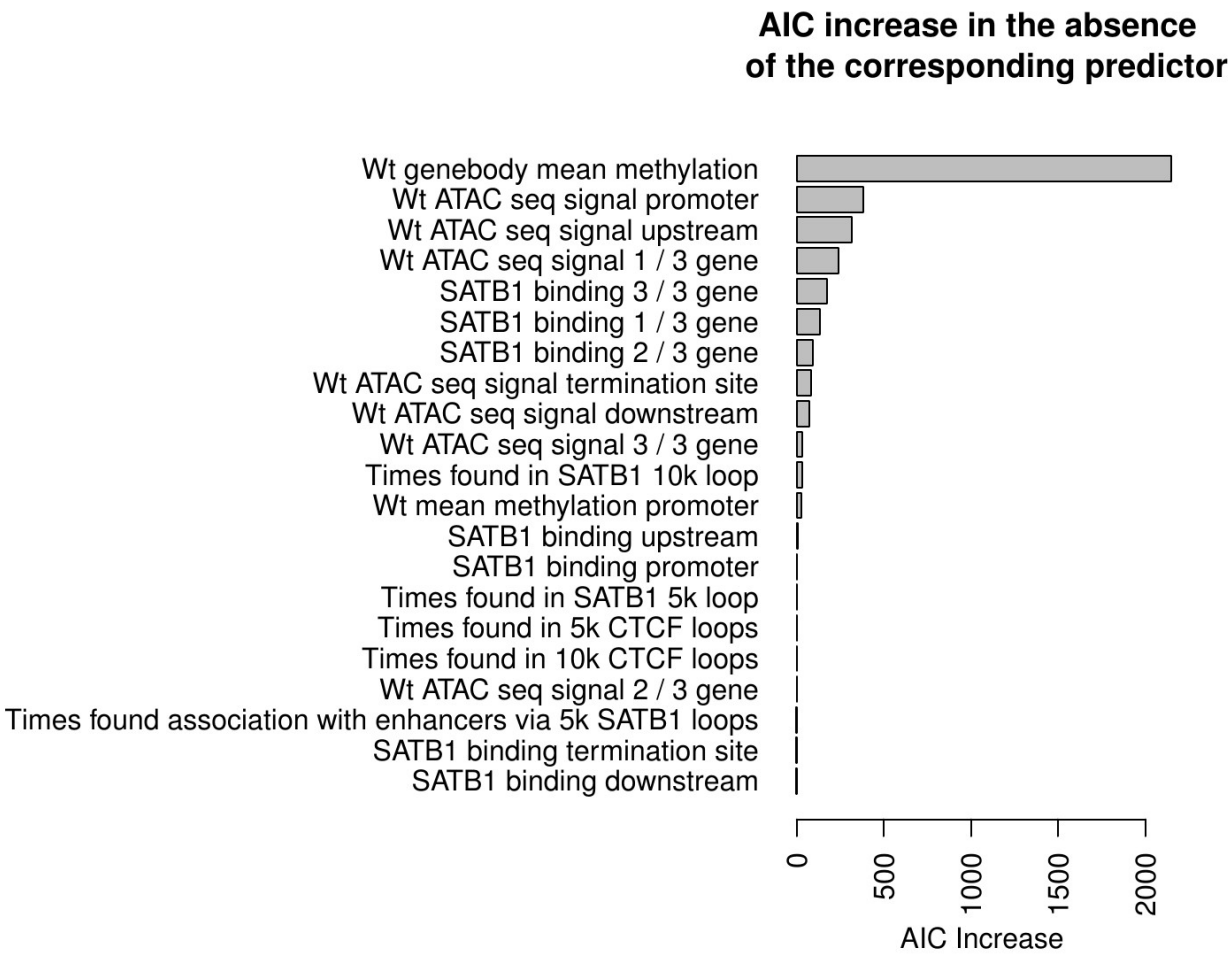
A linear model suggests an activating role for SATB1

The wealth of the information generated by the above datasets can be integrated into a mathematical model : Expression values for a specific condition or changes in expression values between conditions can be modeled using the information from the aforementioned datasets (e.g. ATAC-seq, RRBS). The generated model is not to be used to predict expression values for specific genes. Instead we could observe how specific variables influence the expression of genes through the interpretation of the coefficients for each predictor.

A linear regression model was constructed for the expression values of genes in a Wild-type context or in a *Satb1* cKO context. The expression values of the genes were calculated as follows for each gene:

Mean value of normalized DeSeq2 counts (Wt) / Length of biggest gene transcript

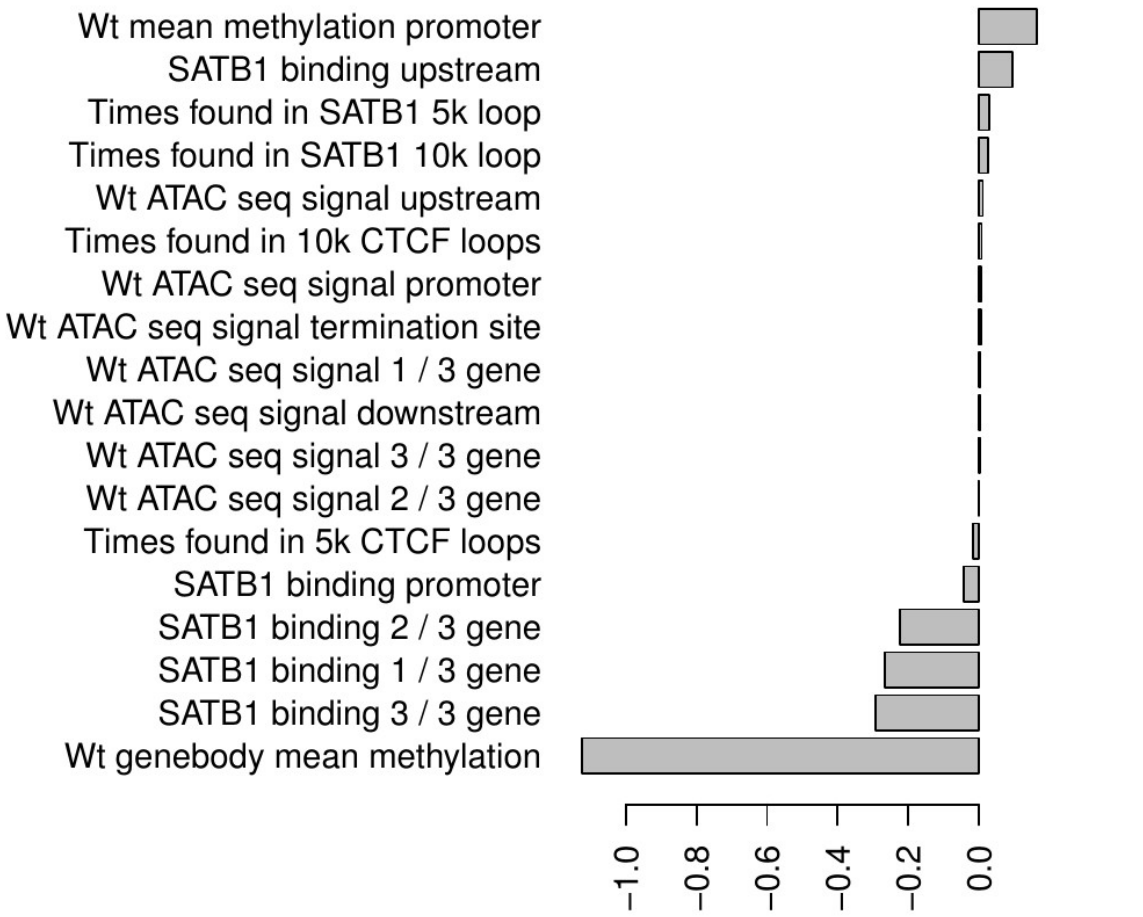
DESeq2 normalizes counts for sequencing depth and RNA composition for each transcript. Thus in order to make intra-sample comparisons possible, the normalized counts for each transcript are divided by the corresponding transcript length. After fitting the model using the predictors stated in the ‘’Materials and Methods’’ section, a backward elimination process and a custom elimination process (see “Materials and Methods”) pointed out the same number of predictors as ‘’important’’.



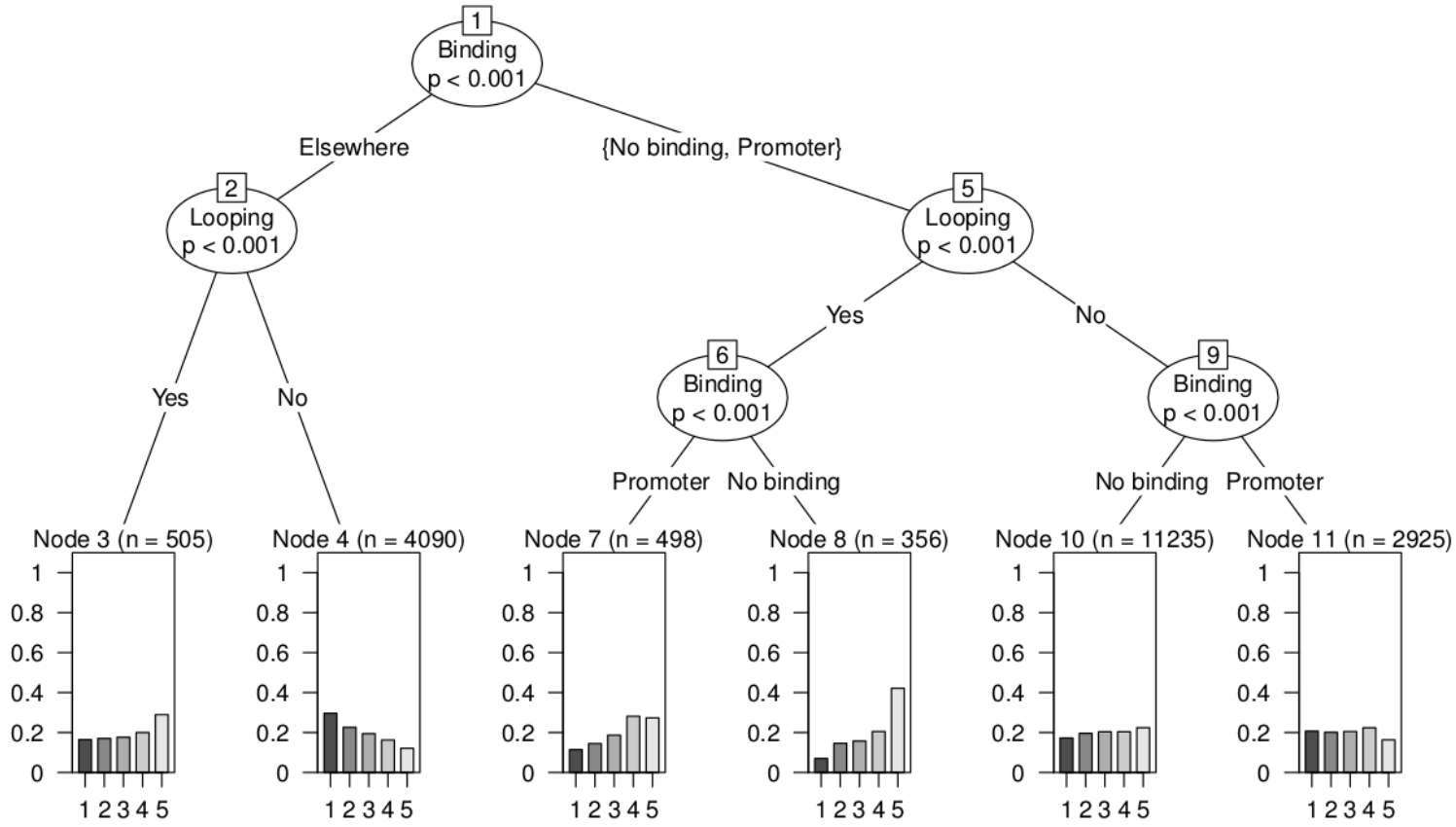
AIC (Akaike Information Criterion) provides a mean for model selection. Here the AIC change of the model is shown when all the predictors are kept intact except one. The y-axis corresponds to the removed predictor, while the x-axis shows how the AIC for the new model is altered. An increase indicates that the predictor was ‘’useful’’ for the model. The last 3 predictors were not ‘’useful’’ for the model.

Apart from the 3 last predictors, every other predictor contributed positively in the accuracy of the model. SATB1 binding at downstream regions does not seem to be significant. Moreover enhancer connections via SATB1 associated loops do not seem to be significant either : This may be due to the fact that these kind of connections are very rare, while the model was build using all the genes exhibiting basal expression levels.

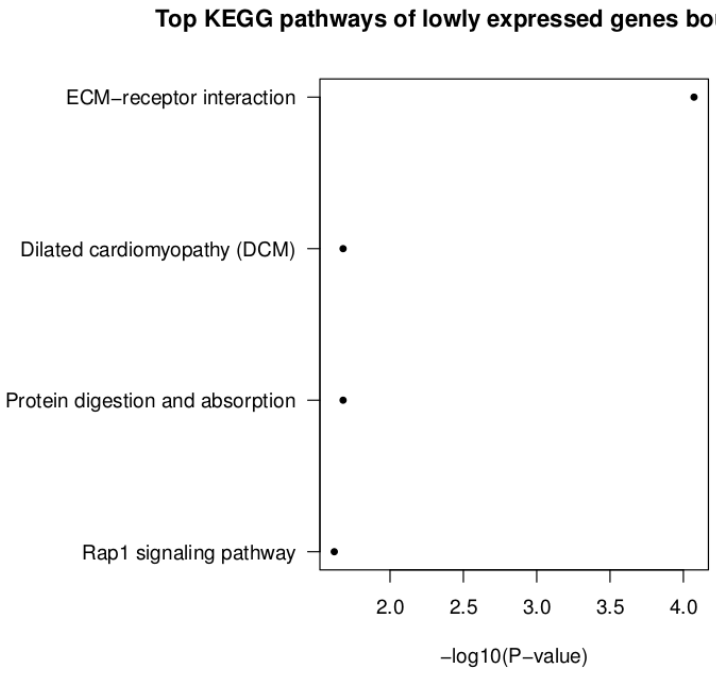
The plot quality metrics showed an excellent behavior (see “Materials and Methods”) after removing some genes that were ‘’influential’’ outliers according to Cook’s distance. The final model has an adjusted R – square of **0.276** , indicating that the model captures a lot of the variability associated with the predicted values. The model coefficients for the important predictors are shown below :



It should be noted that the **sign** of each coefficient is important. It shows whether a predictor is associated with lower (negative) or higher (positive) normalized count values. Genebody methylation and ATAC – seq signal show negative and positive values as expected. SATB1 binding appears to be associated with low expression values. It remains to be seen whether this binding is indicative of a general suppressive function or a characteristic of a pioneer factor. In contrast, looping events associated with SATB1 are indicative of high expression values. This finding was further explored with the use of conditional inference trees :



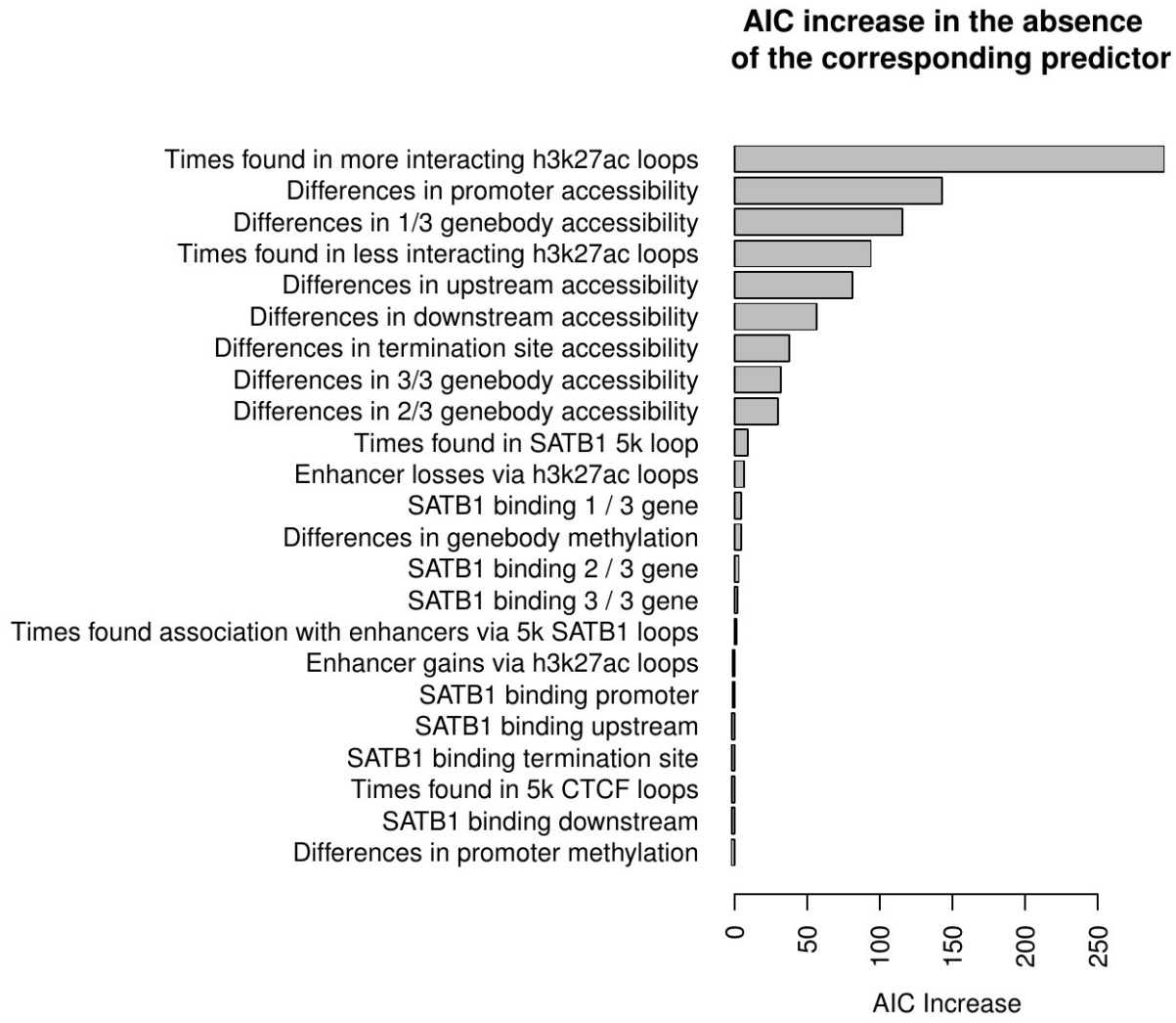
Each bar in the final leaf nodes represent expression levels. The levels were deemed as very low (1 – bottom 20% of expression) , low (2 – 20-40% bracket of expression values) , intermediate (3 – 40-60% bracket of expression values), high (4 – 60-80% bracket of expression values) and very high (5 – 80-100% bracket of expression values). It is evident that looping is associated with high rpkm values, while SATB1 binding at a place different than the promoter is associated with low rpkm values. It remains to be elucidated whether SATB1 binding is **suppressing** expression levels of the genes. The genes that were bound by SATB1 in non-promoter sites that were not found in a loop were isolated. Functional analysis of these genes showed enrichments in the pathways below :



Interestingly, adhesion related genes are the top scorers. A possible action of SATB1 at these genomic loci is suppression. However an interesting hypothesis would be that SATB1 could act as a pioneer factor at these genomic loci, a hypothesis that is in accordance with SATB1's capability of binding nucleosomes. Comparison of SATB1 binding at different tissues could shed light at these questions.

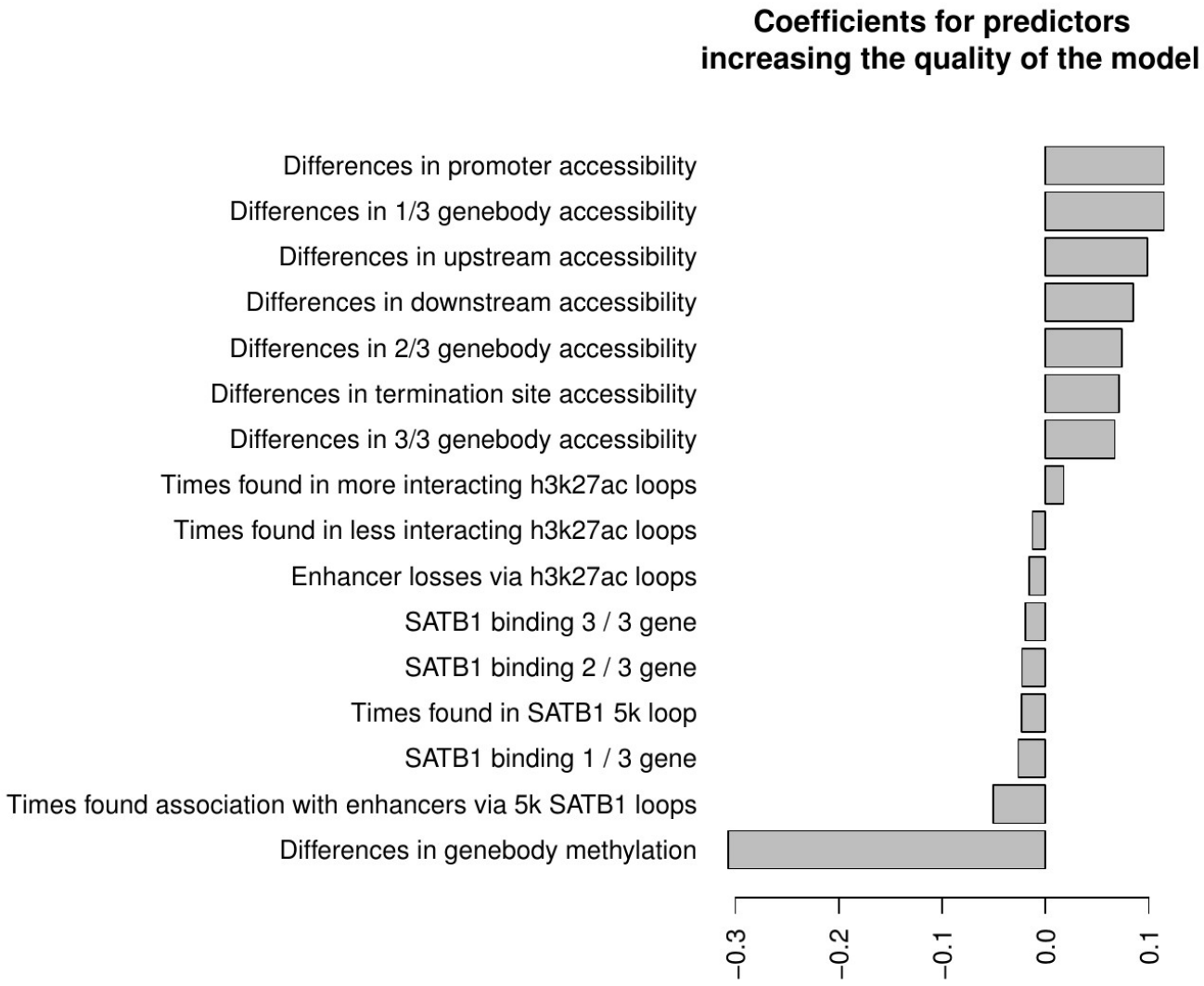
In order to further delineate how SATB1 affects expression another linear model was constructed. This time the predicted values where the logFC changes for each gene, rather than the normalized count values. Moreover a different set of predictors were used (see "Materials and Methods"), since a difference in expression is to be modeled. Quality plots for the model were once again good, but the adjusted R-square of the model was **0.114.**

Once again an AIC plot was constructed for each predictor. The different predictors are shown below:



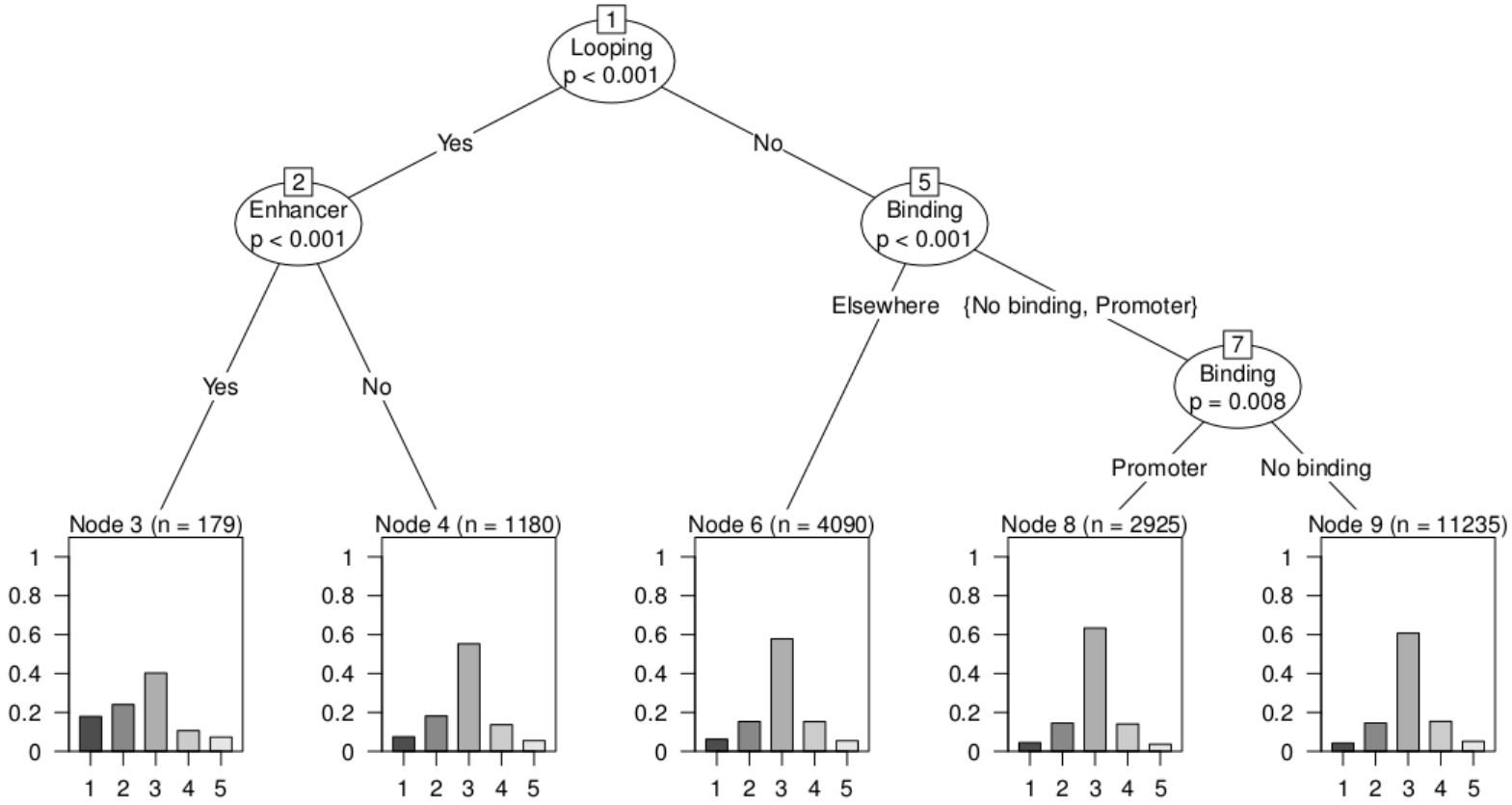
SATB1 binding at various sites along with methylation differences in the promoter regions seem to not contribute to the accuracy of the model. An interesting point is that differences in connectivity via

H3k27ac loops seem to perform very well as predictors. This result points out at the connection between 3D genomic features with transcriptional activity.



The coefficients for the last model reveal some very interesting facts. Once again ATAC-seq values and methylation signal behaves as expected : An increase in accessibility correlates with increased expression values in the *Satb1* cKO cells, while an increase in methylation levels correlates with reduced expression values in the *Satb1* cKO cells. SATB1 binding seems to correlate with reduced expression values : It seems that SATB1 functions mainly as an activator. 5Kb resolution SATB1 associated loops seem to function in the same manner. While the 10kb resolution SATB1 loops seem to exhibit a different trend, the extremely small coefficient for that predictor minimizes its interpretation.

A conditional inference tree was once again constructed for the above dataset :



The bins at the leaf nodes do not contain the same number of genes. 1 denotes the genes that had a dramatic reduction in their expression levels (bottom 5% of the logFC values), 2 denotes the genes that had a reduction in their expression levels (bottom 5-20% of the logFC values), 3 denotes the genes that didn't exhibit differences in expression (20-80% of the logFC values), 4 denotes the genes that showed an increase in expression levels (80-95% of the logFC values) and 5 denotes the genes that showed an extreme increase in their expression levels (top 5% of the logFC values). No evident trends arise from this specific plot, except for the first leaf node. Although the number of genes falling into that category (Included in a SATB1 associated loop – Connected to an enhancer via the loop) are few, an extremely skewed distribution of values towards the left is evident. It seems that the genes connected to enhancers via SATB1 associated loops have their expression values reduced significantly, once again proving that SATB1 can function as an activator.

Discussion

A detailed molecular characterization of the thymus in *Satb1* conditional knockout mice has been carried out. A molecular footprint of the autoimmune phenotype can be identified in the various NGS datasets analyzed.

In normal conditions, it is evident that SATB1 binds a lot of areas throughout the genome. Highly occupied SATB1 areas are close to gene loci. The large amount of binding sites is not restricted to T-cell lineage specific genes. Instead SATB1 binds near to the transcription start site of genes with

‘’housekeeping’’ like functions along with immune related genes. An intriguing finding could point out a pioneer role of SATB1 : Genes exhibiting SATB1 binding throughout their genebodies, but not throughout their transcription start site, were enriched for neural-related functions. Moreover an enriched motif for a neural specific transcription factor was found in the promoter sequences of the genes exhibiting this SATB1 binding pattern. Some of these genes were also classified as differentially expressed genes in a SATB1 knockdown RNA-seq dataset from neural cells. An interesting speculation would be that SATB1 binds a specific set of genes in different cell types irrespectively to the chromatin accessibility of these genomic areas. Although a Chip-seq experiment for SATB1 conducted in other tissues would be needed to support this idea, several additional lines of evidence for such a model are derived from the analyzed datasets. According to the data presented above, it seems that SATB1 can bind sequences occupied by nucleosomes, a finding that was published in a separate study recently (Rajarshi et al., 2018). Finally a lot of SATB1 binding events in genebody regions seem to occur on very lowly expressed genes, according to the conditional inference trees constructed. While this finding could be an indication of a general suppression activity of SATB1, the linear model constructed showed the opposite behavior: In general SATB1 binding to any area of the genbodies correlated with reduced expression levels when SATB1 was deleted.

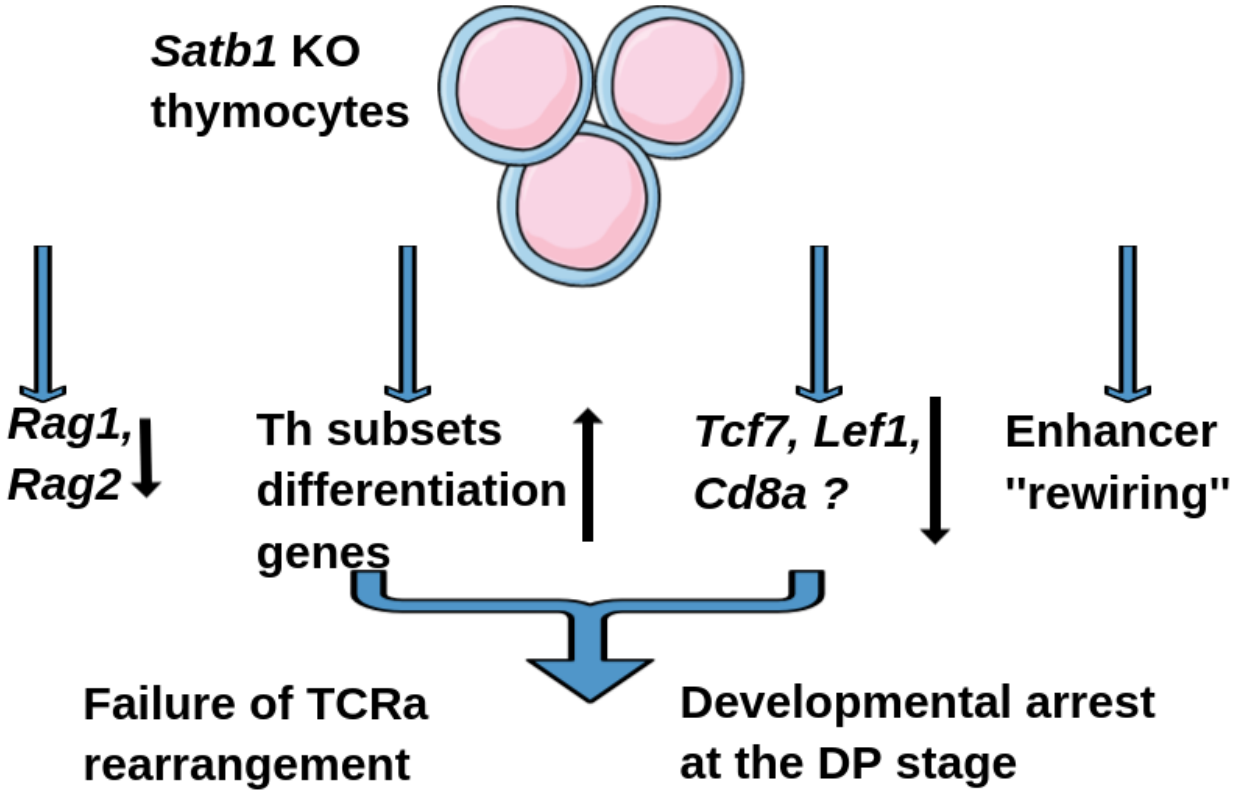
The ablation of SATB1 in thymocytes leads to significant changes in the transcriptional program of these cells. In sorted DP cells it seemed that SATB1 repressed genes were associated with later stages of T-cell differentiation : It also was found that SATB1 bound the promoters of a large fraction of such genes. Activated by SATB1 genes were mostly associated with signaling pathways and were found to be bound by SATB1 in their promoter sites. An interesting difference regarding the underexpressed genes is the SATB1 binding pattern at regions downstream their transcription termination site. This tendency is accompanied by enriched POL2 occupancy and could reflect some peculiar modes of regulation in these genomic regions, like self-interacting short gene loops. The ways that SATB1 influence expression in these gene subsets can vary. Although the association with chromatin remodelers has been described (Krishna et al., 2018), other modes of regulation like establishing contacts with cis-regulatory elements could be possible.

Although SATB1 binds a lot of actively transcribed genes, the expression levels of the majority of these genes didn’t change significantly when SATB1 was deleted. This finding along with the fact that SATB1 tends to co-localize with CTCF probably reflects the fact that SATB1 is a genome organizer. The diffHiC analysis though showed that CTCF is still probably the main organizer involved in the creation of higher order genomic structures. The loops associated with SATB1 showed an extreme overlap with CTCF associated loops. However these common loops were enriched in immune-related genes, while the unique CTCF associated loops didn’t show such enrichments, indicating that SATB1 is found in the bases of context specific gene loops. In general it seems that genes found in SATB1 associated loops are highly expressed, and have their expression levels reduced when SATB1 is missing according to the linear models constructed.

The functional significance of SATB1 associated loops was investigated. SATB1 associated loops in immune-related genes could be reflecting the overall transcriptional activity of these areas. The interplay between 3D-genomic structures and transcription is an extensively studied area and in the majority of cases it is not known whether 3D genome structures drive active transcription or whether the opposite happens (Rowley et al., 2018). Insulation domains could be formed with the aid of SATB1 in such loops: While such insulation appears to occur in a wild-type context, the finding that insulation also persists in the *Satb1* conditional knockout samples, probably indicates that CTCF or other proteins are the culprits for such insulation events. Moreover the loops didn’t seem to contain genes that were expressed in similar levels ruling out the possibility that SATB1 aided in the creation of transcriptional

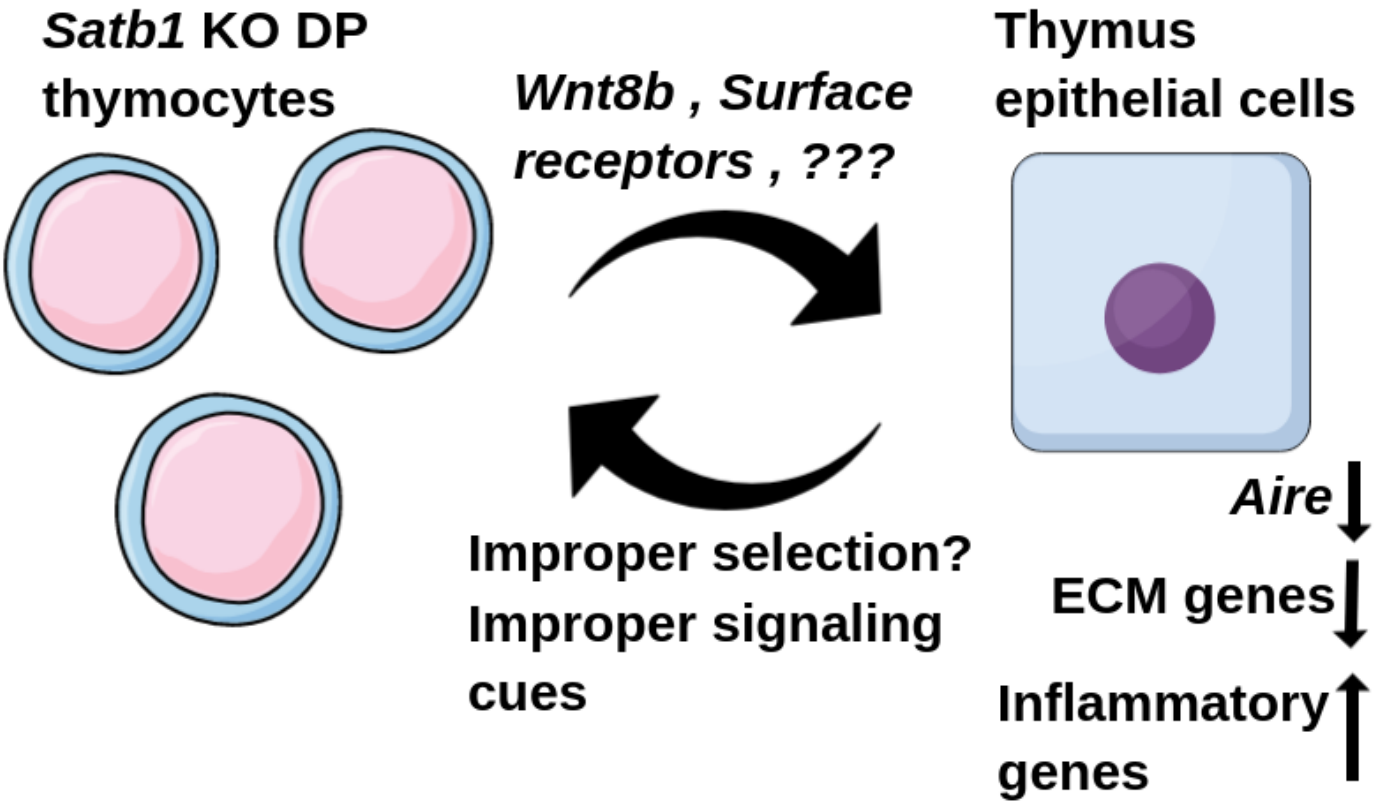
hubs. All of the above pointed out to a transcriptional – driven model, where SATB1 associated loops can be identified mainly due to the fact that highly expressed genes tend to be present in these structures. However the enhancer analysis revealed that a lot of genes that were connected to enhancers via SATB1 associated loops, had their expression values dropped when SATB1 was missing. This finding however is not proving directly that SATB1 mediates these enhancer-genes interactions. The presence of numerous less interacting H3k27ac loops though, that overlap exactly with SATB1 associated loops connected to enhancers, is a convincing finding towards causality.

Some SATB1 associated loops that connected genes to enhancers were also classified as less interacting H3k27ac loops. Genes that were found in such loops were in their vast majority underexpressed in the *Satb1* conditional knockout samples and contained a lot of genes that are crucial for T-cell development. *Rag1* and *Rag2* are such examples : A published paper has already reported the positive regulation of these genes by SATB1 associated loops (Hao et al., 2015). *Tcf7* and *Lef1* were also genes that were identified. The above genes orchestrate extremely important events during T-cell development. *Tcf7* has been characterized as a pioneer factor, opening a lot of T-cell specific genomic areas throughout their development and the combined action of both of these genes is needed for the transition to the CD4+ single positive stage by regulating the expression levels of the *Thpok* gene (Johnson et al., 2018)(Farrah et al., 2014). The developmental block of DP thymocytes in the absence of SATB1 is already reported (Alvarez et al., 2000) and was mainly attributed to the reduced expression levels of *Rag1* and *Rag2* (Hao et al., 2015). *Tcf7* regulates the *Thpok* axis needed for proper lineage specification to CD4+ single positive cells and *Thpok* is reduced ~8 fold in our RNA-seq dataset : Based on the provided evidence, it seems that SATB1 directly regulated the expression levels of *Tcf7* and this event could play a part in the developmental block of the DP thymocytes in the absence of SATB1.



It also seems that the depletion of SATB1 from DP thymocytes affect other cells of the thymus. A lot of differentially expressed genes were identified in the thymus RNA-seq dataset that were not even expressed in the dataset derived from sorted DP cells. Although this effect could be attributed to the different sequencing depth between experiments, a lot of the genes isolated by this procedure are known to be only expressed in thymus epithelial cells : The best example of such a gene is *Aire*, a gene needed to express a wide variety of antigens in thymic epithelial cells (Giraud et al., 2012). A lot of adhesion related genes expressed in the thymus epithelial cells are underexpressed. In contrast genes that are related to inflammatory responses and cytokine receptor – cytokines contacts are overexpressed. Finally differences in the methylation status of the *Ugt* family genes can be observed: Once again this gene family is expressed in thymic epithelial cells. These results point out to a molecularly disrupted thymus environment and could provide the basis for the phenotypic observations of the thymus, like the shrinkage of it.

The big question remains : How does the ablation of *Satb1* in thymocytes drive such extreme changes in the homeostasis of the epithelial cells? It is obvious that the crosstalk between the different cell subsets is the culprit for these changes. Distinct signaling pathways could be responsible for these changes and one of those could be the *Wnt* pathway. As it seems, a *Wnt* ligand, *Wnt8b*, is highly overexpressed in thymocytes after the deletion of *Satb1*. It is known that thymic epithelial cells respond to the *Wnt* pathway and excessive *Wnt* signaling can disrupt their normal functions (Swann et al., 2017). Another process which could lead the thymic epithelial cells to ‘’malfunction’’ is the cell contact between those cells and DP thymocytes. A lot of surface receptors found in both cell subsets are needed for proper *Aire* expression for example. If the concentration of these receptors in the surface of DP cells is altered, one could expect changes in the physiology of the epithelial cells (Noëlla Lopes et al., 2015).



The consequences of the differences in the thymus environment could hold the key in understanding autoimmunity. Perhaps the autoimmune phenotype should not be attributed solely to T-cell intrinsic defects associated with the ablation of the *Satb1* gene. Maybe the differences in the thymus structure lead to improper T-cell selection, allowing a lot of autoreactive (and also defective) T-cells in the periphery. Further study could shed light for this model and possibly determine a major cause of autoimmunity in this experimental system.

Materials and Methods

Generation of conditional knockout *Satb1* mice

A mouse line with loxP sites flanking the third *Satb1* exon was crossed with mice expressing the CRE recombinase under the control of the CD4 promoter.

RNA-seq analysis of sorted DP cells

Bowtie2 was used to align the generated fastq files with standard parameters (Langmead et al., 2012). The mm10 genome assembly was used. Differentially expressed genes were extracted with the aid of an R package named metaseqR (Tsuyuzaki K et al., 2019). Read counts were quantified on the gene level. Genes exhibiting a change of absolute(log2FC)>=0.5849 along with a p-value<=0.05 were classified as differentially expressed genes.

Gene files

2 different gene files were used for analyses corresponding to general tendencies across genes. The first gene file was retrieved by the UCSC table browser (Karolchik et al., 2004) by setting the following parameters:

Mouse → mm10 → Genes and gene predictions → NCBI Refseq → UCSC RefSeq

The final gene list was generated by keeping the unique genes only : The coordinates of the biggest transcript for each gene (Transcription start site – Transcription end site) were kept as the gene's coordinates. Note that in some cases the transition to the gene coordinates of the mm9 assembly was needed. In these cases the liftOver tool of UCSC was used.

Analyses associated with the SATB1 Chip-seq dataset and the RNA-seq dataset of sorted DP cells were carried out by using the above gene list.

The second gene file was retrieved by the UCSC table browser by setting the following parameters:

Mouse → mm10 → Genes and gene predictions → All GENCODE VM18 → Comprehensive

The final gene list was generated again as described above. With the exception of the analyses that were carried out with the previous gene file, all the other analyses were carried out with this dataset.

SATB1 Chip-seq

A publicly available Chip-seq dataset was downloaded and used for various analyses. The accession number was GSM1617950. The Chip-seq analysis pipeline is described by the team that published the dataset.

Functional analysis of gene lists

For all the functional analysis performed, gProfileR (Karolchik et al., 2019) was used. The plots depicting enriched BP terms and KEGG pathways are generated by picking the top 20 pathways/terms with the lowest p-values. In case where enriched pathways/terms were less than 20, all of the enriched pathways/terms are depicted.

Average gene profiles

For each region of interest, an empty vector of the same length was constructed. Each position of the vector corresponded to a base. For example if an average gene profile of areas upstream of the transcription start site was to be calculated (e.g. -4 kb upstream of the TSS, until the TSS), then a vector with 4001 positions would be constructed. The first position of the vector would correspond to the position -4kb, while the last position would correspond to the transcription start site. Each position was filled with a score indicating either the total number of times a peak was found to overlap with this base, either with the sum of the reads found overlapping with this base. The read density for such plots was only used in order to construct the POL2 related average gene plots and the accessibility density scores across SATB1 peaks. Scores for each average gene profile were normalized by the number of ‘’objects’’ of interest. For example when calculating the average gene profiles of SATB1 binding at the promoters of underexpressed genes, the final average gene scores were divided by the total number of underexpressed genes.

In order to plot the generated average gene profiles for each area of interest, the runmean function in R was employed. Specifically the average gene profiles were smoothed with a window of 100 and were afterwards plotted.

SATB1 binding heatmap

For the generation of the SATB1 binding heatmap each gene was binned in 10 equally sized bins. The bin size for each different gene was variable and was analogous of its length. If a SATB1 peak overlapped with a bin, the corresponding bin was marked as an ‘’1’’ bin. If no overlap existed the bin was marked as a ‘’0’’ bin. The same procedure was followed for regions upstream and downstream of the gene coordinates for each gene: The difference here is that fix sized bins were used to describe the upstream and downstream areas (bins of 400bp were used, for a total of 10 bins). The final heatmap was plotted with the aid of the gplots package in R. The clustering method used was the “Ward.D2” method.

HiChIP analysis

Since the HiChIP experiment was carried out in female thymi (while the SATB1 Chip-seq dataset was generated by male mouse thymi), a different set of peaks was called using an old HiChIP dataset. The old experiment contained a lot of reads that cannot be used to call contact frequencies between genomic regions (e.g. self-ligation reads). However these reads should be enriched for SATB1 binding sites, since they were isolated after an immunoprecipitation step with an antibody specific for SATB1. Thus

MACS2 (Zhang et al., 2008) was applied on these kind of reads in order to generate a new set of peaks. MACS2 was run with standard parameters except for the following tweaks :
--nomodel --extsize 147
Furthermore, in order to get a set of ‘’confident’’ binding sites, peaks that showed a >2.5 enrichment score were only kept. The final peaks were filtered for mitochondrial peaks, unknown chromosomes and were merged using the bedtools (Aaron R et al., 2010) merge command.

Using FitHiChIP, a set of loops was isolated. FitHiChIP was run with standard parameters and the option “Peak to All”. Moreover a wrapper FitHiChIP script was used in order to isolate differentially interacting H3K27ac loops. The script operates on count-based methods (edgeR) to identify such cases (Sourya Bhattacharyya et al. , 2019).

Isolation of SATB1 binding hotspots

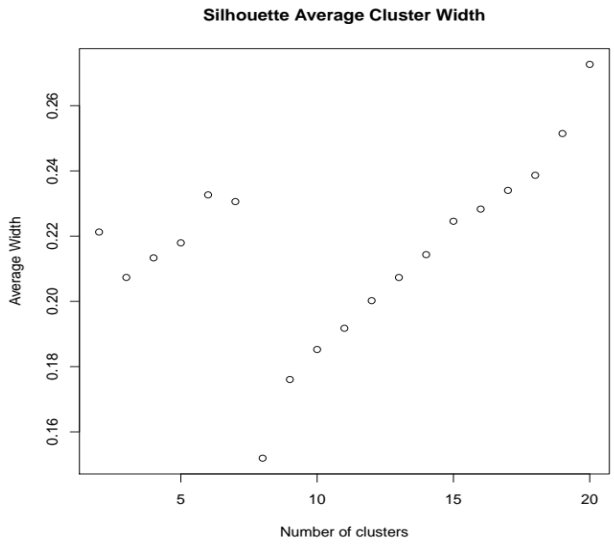
To reveal places in the genome that showed great SATB1 occupancy (based on the publicly available peaks), a sliding window approach was employed. Beginning from the start of each chromosome, a window of 100kb was slided up until the end of the chromosome. The window was moved 10kb each time. For each location of the window, the fraction of the bases occupied by SATB1 peaks was calculated. Finally a background model was constructed in order to evaluate whether a window had a high SATB1 occupancy score. A window was selected randomly and was kept constant. The Krangel’s peaks were shuffled with the bedtools shuffle command and each time the fraction of bases of the constant window that were found occupied by SATB1 was retrieved. Using these 10000 fraction values, a p-value was calculated for each window i:

p-value (i) = (Number of fraction values larger than the fraction of the window i / 10000)

The p-values where then corrected using the Benjamini-Hochberg correction in order to take into account the mulitple comparisons performed. A cutoff of p-value <=0.05 and corrected – pvalue <=0.05 was applied in order to extract the final hotspots. Finally the bedtools merge command was used in order to ‘’melt’’ overlapping windows.

Deciding the number of clusters in heatmaps

The number of clusters of genes, according to the binding pattern of SATB1, was not decided randomly. The R silhouette function was used to determine the optimal number of clusters. More specifically clustering was performed for various number of possible groups and the average silhouette width was calculated for each number.



Clustering was performed for a final of 6 clusters. As it seems this number achieves a relatively small amount of clusters alongside with a local maximum of the average width silhouette metric.

Motif analysis

In order to determine possible enriched motifs in sequences of interest, the MEME-suite (Timothy et al., 2009) was used. MEME-ChIP was used for all kind of motif analysis.

RNA-seq dataset of *Satb1* knockdown in neurons

Gene count files were downloaded from GSE70133. EdgeR (Robinson et al., 2010) was used at these count files with standard parameters in order to determine differentially expressed genes. The same cutoffs were used in order to determine differentially expressed genes.

Permutation analyses

In order to construct a null statistical model for various experiments (e.g. determination of common genes between two gene subsets, estimation of expected peak overlaps) permutation analyses were carried out. In cases of overlaps between two files with genomic coordinates, the coordinates of one file were shuffled 1000 times with the bedtools shuffle command. Afterwards, the new overlap for each iteration was calculated. The mean overlap count was used to determine enrichments. Finally a p-value was calculated as follows :

P-value for comparison i = (Number of times an overlap was higher than the actual overlap /1000)

The same logic was applied for overlaps between gene lists, with the exception of drawing out random genes instead of shuffling coordinates.

ATAC-seq analysis

The ATAC-seq fastq files for each condition (Wt mice, *Satb1 cKO* mice) were used as the input of the R package esATAC (Wei et al., 2018). This R package is an integrative pipeline carrying out all the necessary steps needed for the analysis of an ATAC-seq dataset : Quality controls of the fastq files, trimming of adapter, mapping, generation of big wig files, peak calling, footprint analysis. All the samples passed the quality plots and metrics calculated.

Accessibility scores within SATB1 peaks

The accessibility scores calculated around SATB1-centered peaks were derived from the number of reads falling withing these centered peaks (at each position). The plots were constructed in the same way the average gene profile plots were calculated. The three samples of each condition were used to generate the final scores. An example is given below :

Wt (i) → Score of the -1000 position : Sum of ATAC-seq reads overlapping with this position / Correction term i

The final score for the Wt samples for the position -1000 is the mean score of all the above Wt(i) scores. The correction term applied for each sample normalizes for the different sequencing depths across the different samples.

Identification of differentially accessible regions

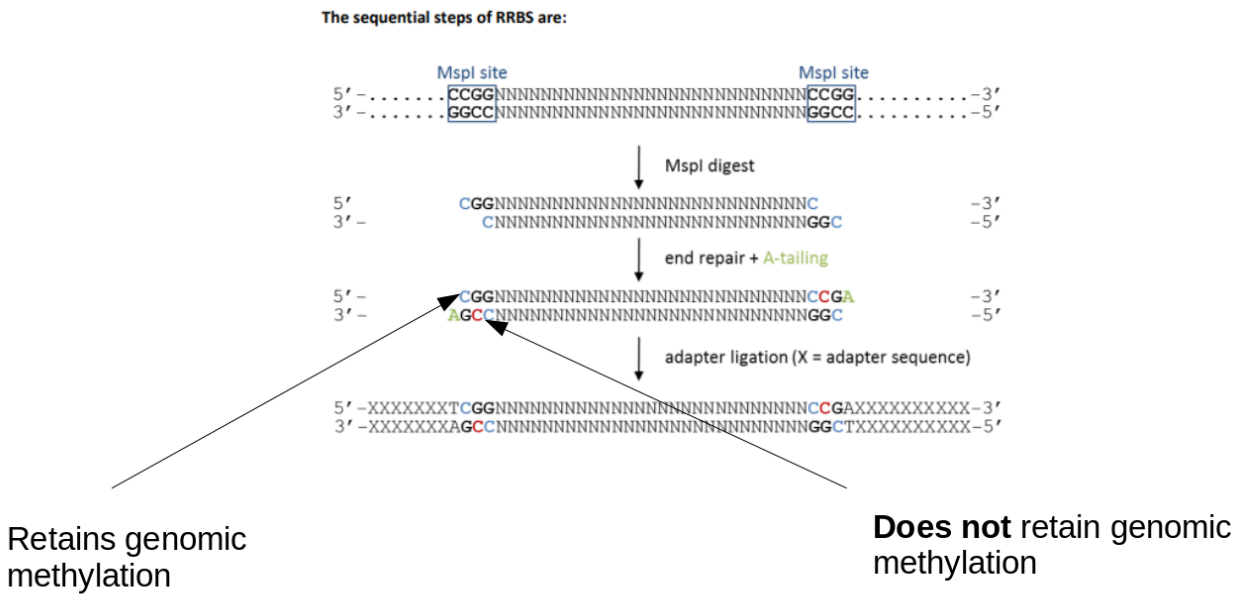
In order to determine regions showing statistically significant differences in accessibility, edgeR was used. The whole mouse genome was binned in bins of 1000bp. The ATAC-seq counts for each bin was calculated for each ATAC-seq sample. These counts were provided to edgeR, which was run with the standard parameters, in order to detect differentially accessible regions. A cutoff of abs(logFC)>=1 alongside a p-value <=0.05 was used to determine these regions. The final regions were calculated after merging adjacent bins that showed the same behavior (less accessible or more accessible).

KEGG pathway plots

Pathway plots were generated with the aid of the R package ‘’Pathview’’ (Luo et al., 2013).

Analysis of RRBS datasets

TrimGalore (https://github.com/FelixKrueger/TrimGalore) was used to trim the ends of reads that showed overlap (even partial) with known Illumina Adapter sequences. This harsh trimming was performed since the adapters can contaminate an RRBS experiment as shown in the below picture.



The trimmed reads were aligned to the mm10 genome with the aid of Bismark (Krueger et al., 2011), an aligner specifically designed for such kind of datasets. Bismark was run with standard parameters except that one mismatch was allowed between a read and a genomic position. Bismark also reports the methylation counts (times a cytosine was found methylated and times the same cytosine was found non-methylated) for each cytosine in the genome. An R package, DSS (Feng et al., 2014), was used to extract differentially methylated regions using these counts. The following parameters were used to extract the differentially methylated regions :

Minimum region length of 50bp, p-value threshold of 0.00001, minimum CG of 3, percentage of significant CpGs of 50%, merging distance of 100 bases

Nucleosome profiles of Wt and *Satb1* conditional knockout thymi

The SATB1 peaks and the CTCF peaks coordinates were used as the input for NucleoATAC (Schep et al., 2015), an algorithm that infers nucleosome positioning by ATAC-seq data. NucleoATAC was run with standard parameters on the merged for each condition ATAC-seq bam files.

Linear model construction

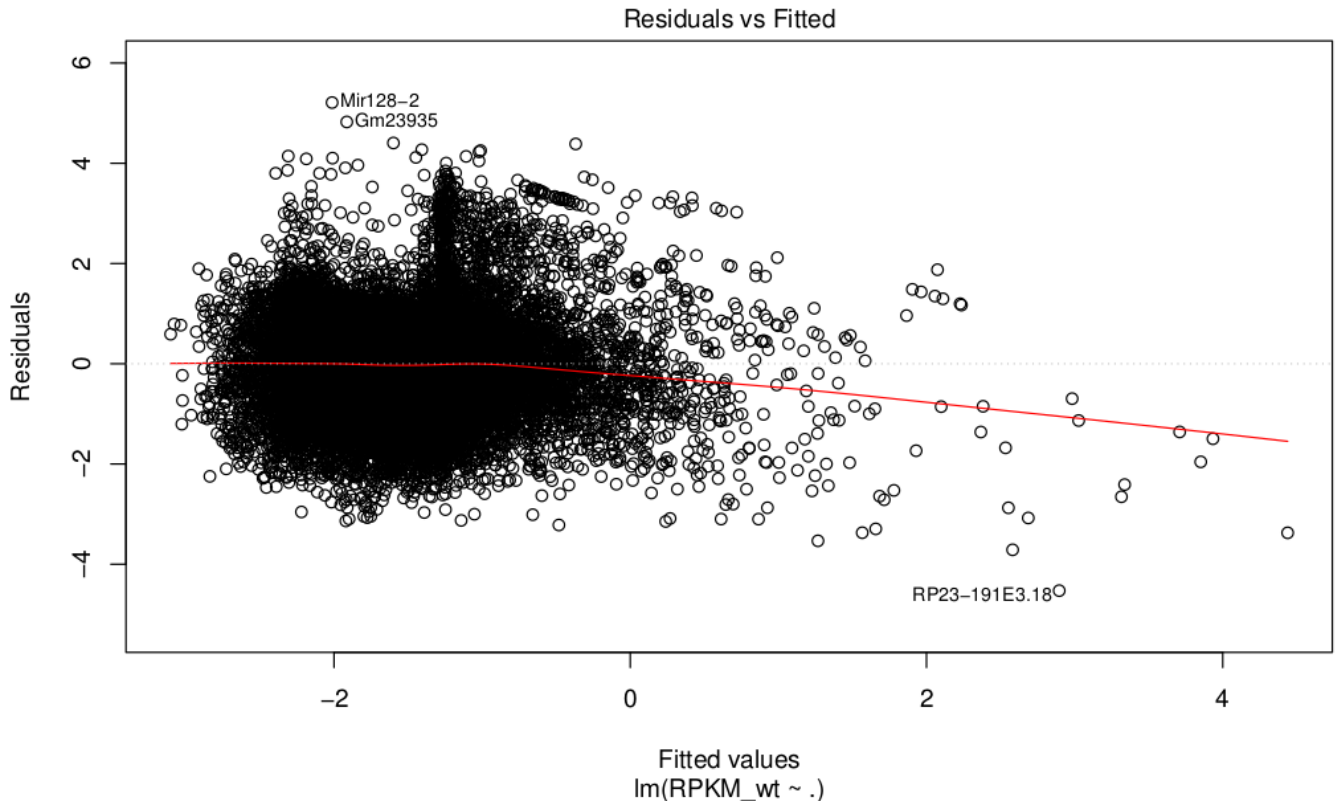
The standard R command *lm* was used to create a linear model. Some clarifications regarding the values used for specific variables are given below.

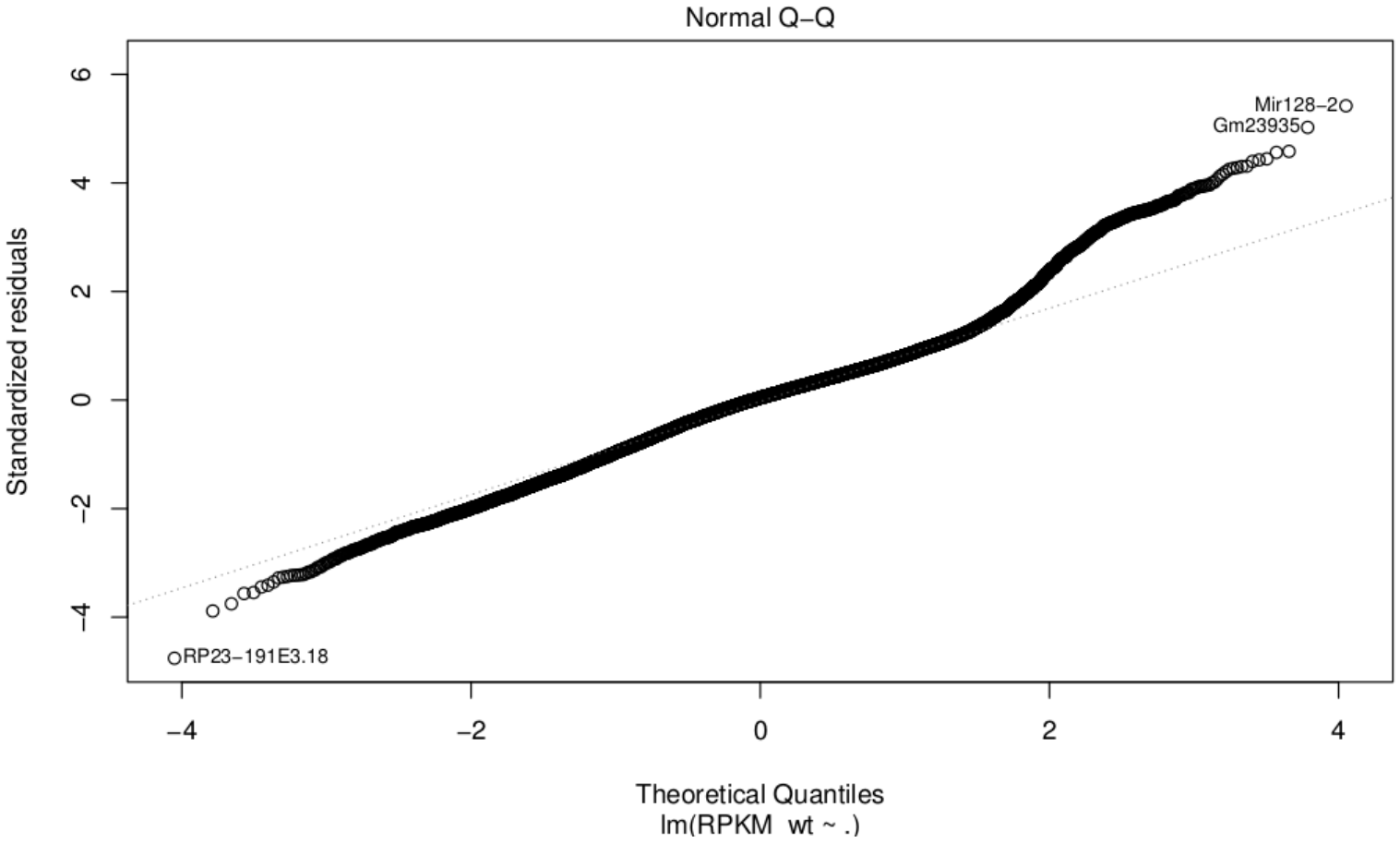
Wild-type thymi RPKM values as the response variable :

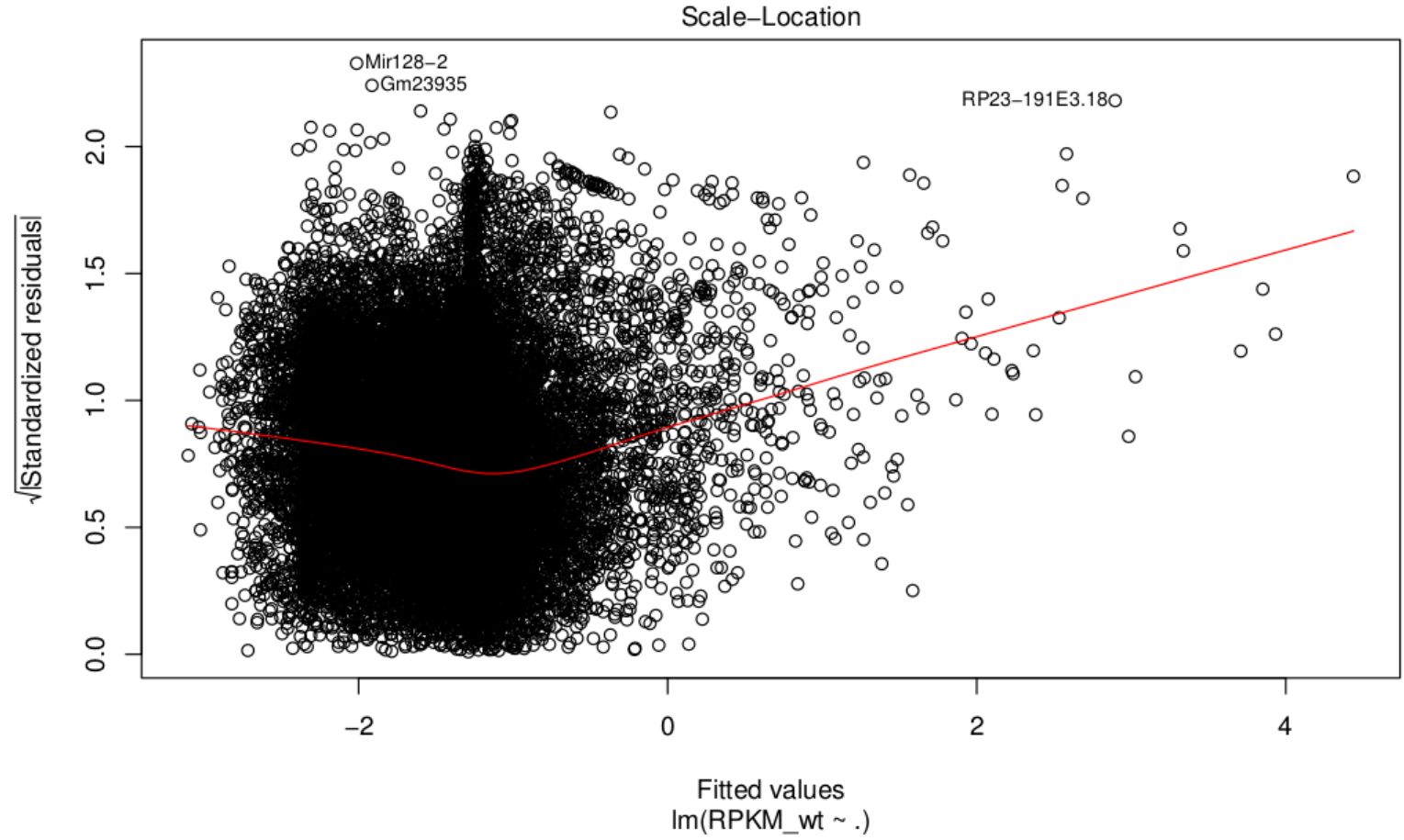
ATAC-seq signal values for each gene : Each gene was binned in 3 bins. An accessibility score was calculated for each bin in a manner similar as described in the "Accessibility scores within SATB1 peaks" section. Moreover the same score was calculated for two extra bins upstream of the transcription start site of each gene(upstream region : -4kb to -2kb and promoter region : -2kb to TSS) and for two extra bins downstream of the transcription termination site (TTS to +2kb and +2kb to +4kb).

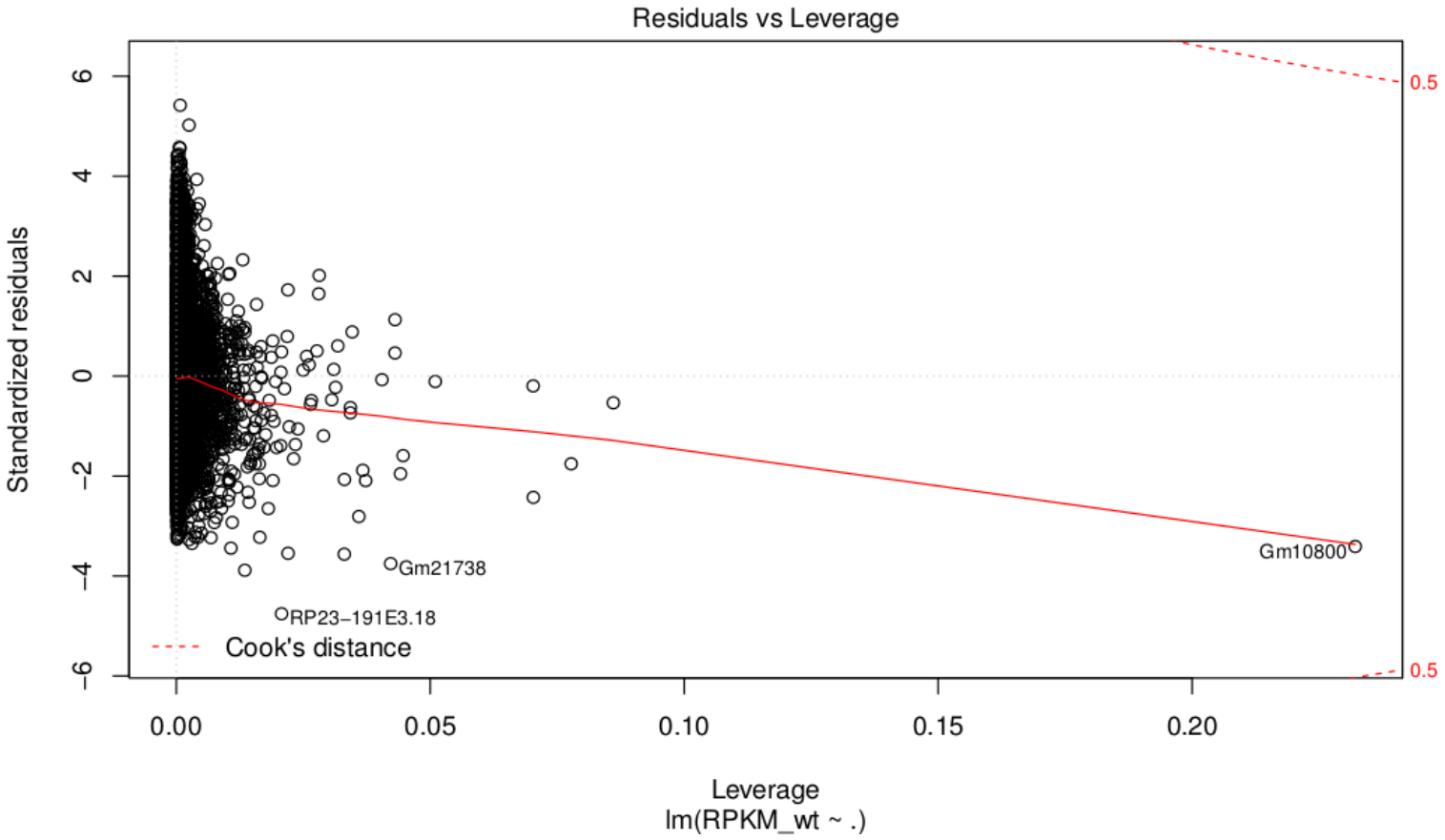
Methylation values for each gene : The mean methylation score for each genebody was extracted by calculating the average methylation proportion value of all the candidate for methylation cytosines found in each genebody. The same was done for the promoter regions (-1kb – TSS for each gene).

Quality plots for wild-type thymi RPKM values :

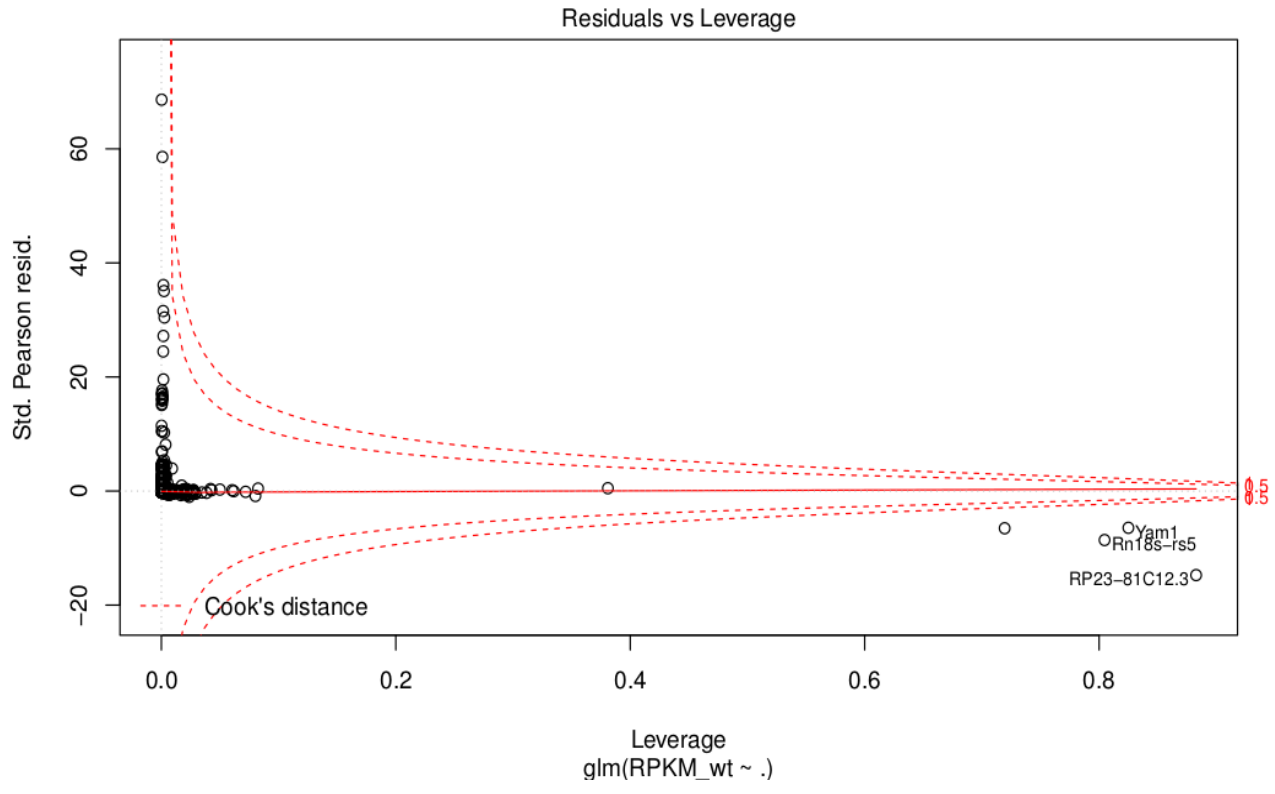



Normal Q–Q
Standardized residuals
Theoretical Quantiles
lm(RPKM wt ~ .)
Mir128–2
Gm23935
RP23–191E3.18


Scale–Location
√|Standardized residuals|
Fitted values
lm(RPKM_wt ~ .)
Mir128–2
Gm23935
RP23–191E3.18



It has to be noted that certain genes had to been removed due to their classification as ‘’influential outliers’’ according to the Cook’s distance quality metric. The ‘’Residuals vs Leverage’’ plot of the old model containing these genes is the following:



logFC changes as the response variable:

ATAC-seq signal differences : To quantify such signal differences between the *Satb1* conditional knockout thymi and the wild-type thymi, the below score was employed for each bin of each gene.

$$\log10\left(\frac{\text{SKO signal} + 0.01}{\text{Wt signal} + 0.01}\right) * \log2(\text{Total signal} + 1)$$

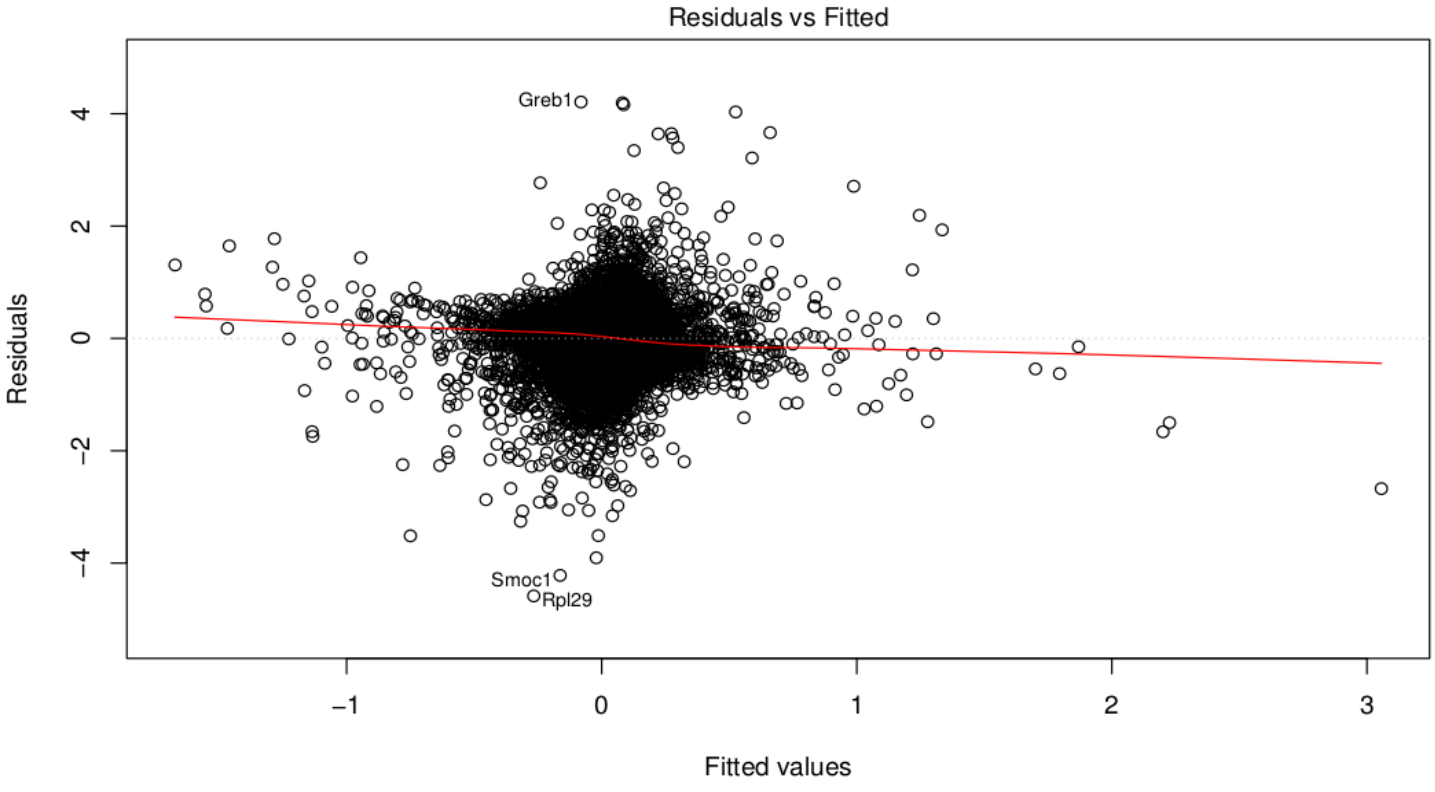
The first term quantifies the difference between conditions. Log transformation adds a positive or negative sign to the difference

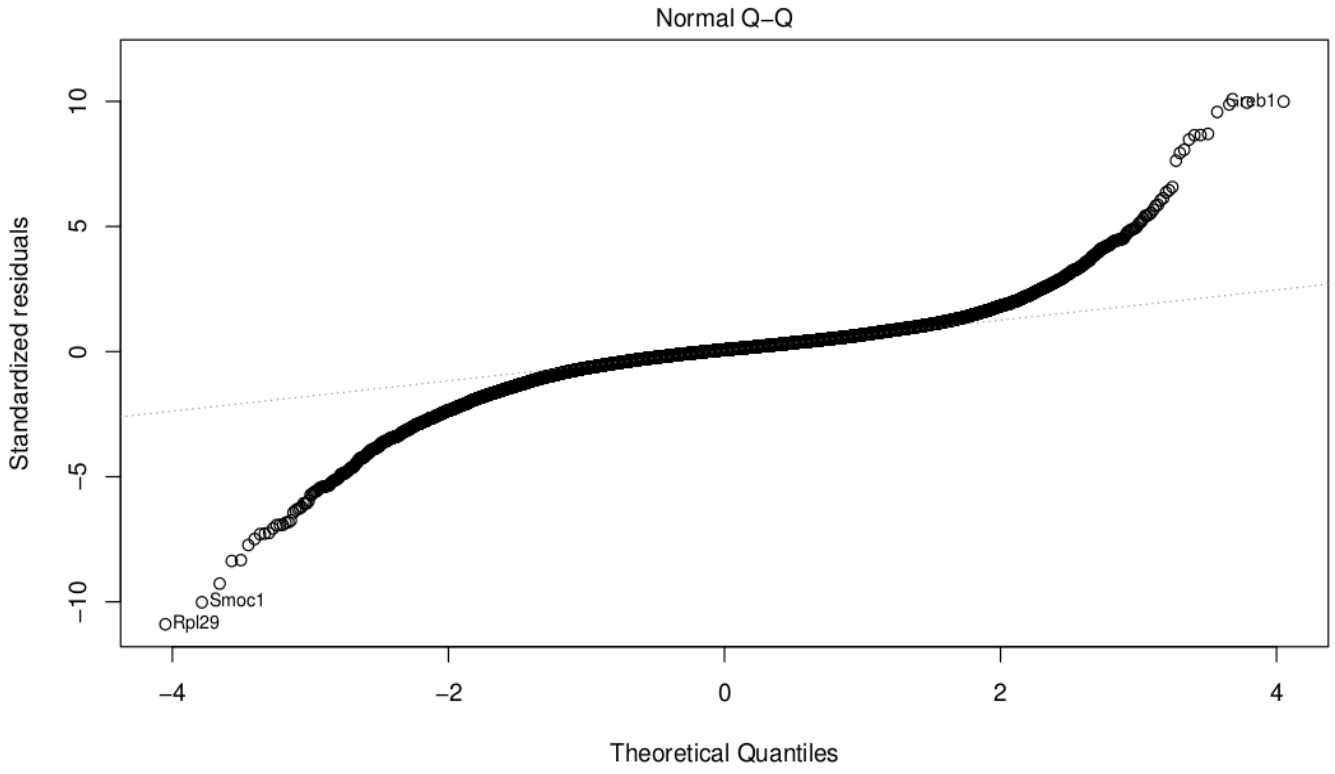
Shrinkage term. Adding one to the score leads to positive values only. If the total ATAC-seq signal is greater than 1, then this parameter increases the overall score.

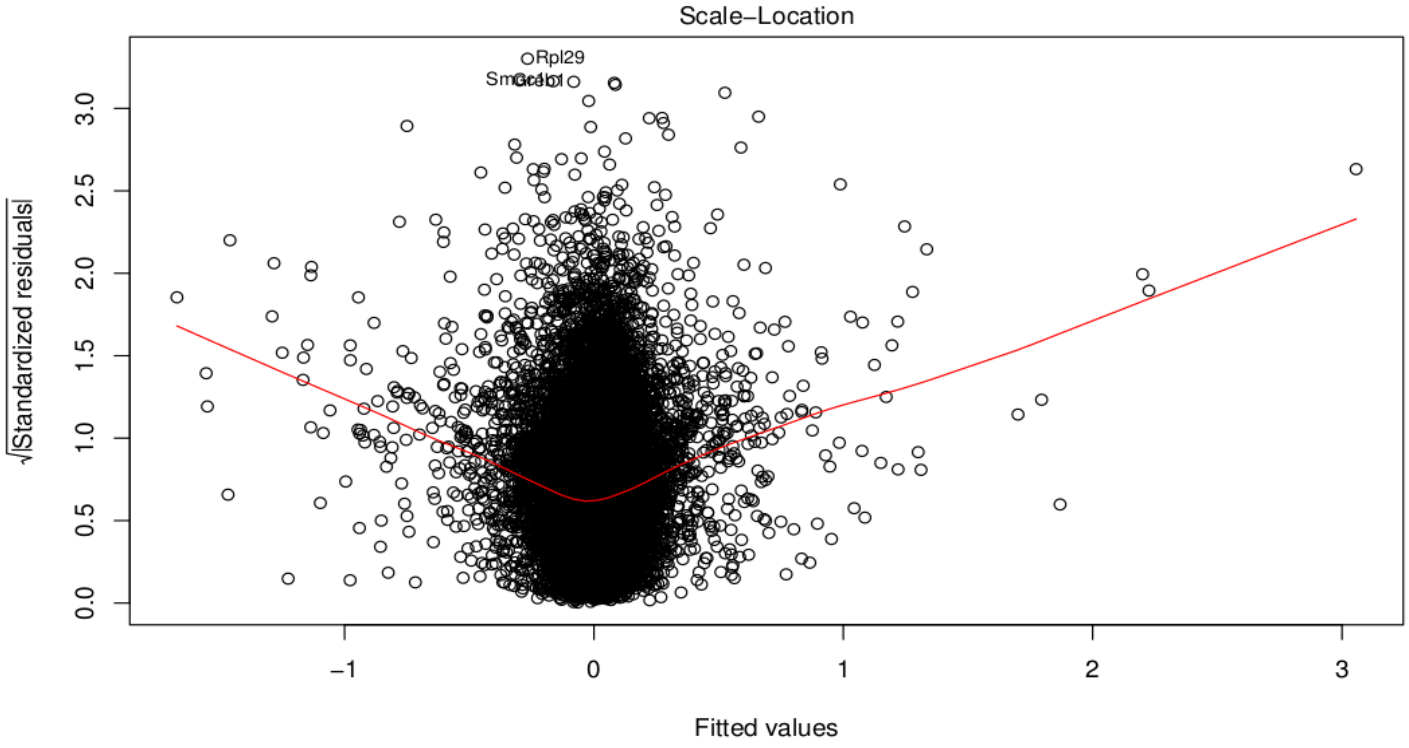
Differences in methylation proportions : To quantify such differences a simple score was calculated.

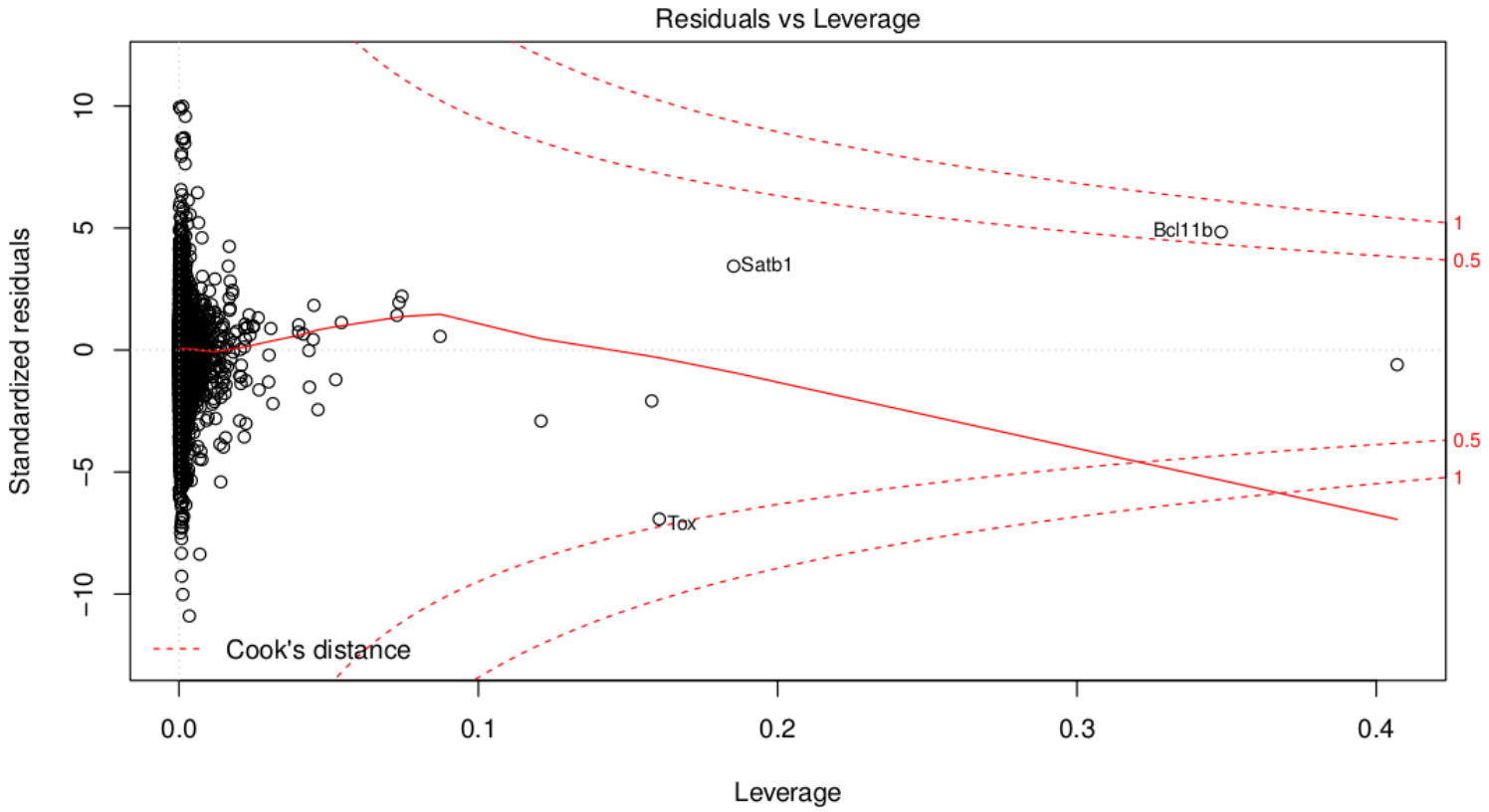
Methylation difference (i) = Methylation proportion *Satb1* cKO – Methylation proportion Wt

Quality plots for the model :




Normal Q–Q
Standardized residuals
Theoretical Quantiles
Smoc1
Rpl29
10
5
0
−5
−10
−4
−2
0
2
4


Scale–Location
√|Standardized residuals|
Fitted values
Rpl29
0.0
0.5
1.0
1.5
2.0
2.5
3.0
−1
0
1
2
3



References

Molecular biology of the cell, 4th edition. Bruce Alberts,Alexander Johnson,Julian Lewis,Martin Raff,Keith Roberts, andPeter Walter. Garland Science, 2002.

Mechanisms of thymus organogenesis and morphogenesis. Julie Gordon and Nancy R.Marley. 2011 Development, Sep 15; 138(18): 3865–3878.

Mechanisms of T-cell development and transformation. Ute Koch and Freddy Radtke. Annual review of cell developmental biology, Volume 27 , 2011.

Mechanisms of T-cell receptor rearrangement. Michael Krangel. Current Opinion Immunology, 2009 Apr; 21(2): 133–139.

A tissue-specific MAR/SAR DNA-binding protein with unusual binding site recognition. Dickinson LA,Joh T,Kohwi Y,Kohwi-Shigematsu T. Cell, 1992 Aug 21;70(4):631-45.

The MAR-binding protein SATB1 orchestrates temporal and spatial expression of multiple genes during T-cell development. John D. Alvarez, Dag H. Yasui, Hiroyuki Niida, Tadashi Joh, Dennis Y. Loh and Terumi Kohwi-Shigematsu. Genes Dev. 2000 Mar 1; 14(5): 521–535.

Functional relevance of SATB1 in immune regulation and tumorigenesis. Krishna P. Sunkara, Gaurav Gupta, Philip M.Hansbroc, Kamal Dua, MaryBebawy. Biomedicine & Pharmacotherapy , Volume 104, August 2018, Pages 87-93
SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. Cai S,Lee CC,Kohwi-Shigematsu T. Nat Genet. 2006 Nov;38(11):1278-88. Epub 2006 Oct 22.

Phosphorylation of SATB1, a global gene regulator, acts as a molecular switch regulating its transcriptional activity in vivo. P. Pava Kumar,P.K.Purbey,C.K.Sinha,D.Notani,A.Limaye,R.S.Jayani,S. Galande. Mol. Cell,22(2)(2006), pp.231-243.

Satb1 ablation alters temporal expression of immediate early genes and reduces dendritic spine density during postnatal brain development. Balamotis MA,Tamberg N,Woo YJ,Li J,Davy B,Kohwi-Shigematsu T,Kohwi Y. Mol Cell Biol.2012 Jan;32(2):333-47. doi: 10.1128/MCB.05917-11. Epub 2011 Nov 7.

SATB1 Defines the Developmental Context for Gene Silencing by Xist in Lymphoma and Embryonic Cells. Ruben Agrelo, Abdallah Souabni, Maria Novatchkova, Christian Haslinger, Martin Leeb, Vukoslav Komnenovic, Hiroyuki Kishimoto, Lionel Gresh, Terumi Kohwi-Shigematsu, Lukas Kenner, and Anton Wutz. Dev Cell. 2009 Apr; 16(4): 507–516.

SATB1 is Correlated with Progression and Metastasis of Breast Cancers: A Meta-Analysis. Pan Z, Jing W, He K, Zhang L, Long X. Cell Physiol Biochem. 2016;38(5):1975-83. doi: 10.1159/000445558

An anti-silencer- and SATB1-dependent chromatin hub regulates Rag1 and Rag2 gene expression during thymocyte development. Hao B, Naik AK , Watanabe A, Tanaka H, Chen L, Richards HW, Kondo M, Taniuchi I, Kohwi Y, Kohwi-Shigematsu T, Krangel MS. J Exp Med. 2015 May 4;212(5):809-24. doi: 10.1084/jem.20142207. Epub 2015 Apr 6.

Engineering Mouse Chromosomes with Cre-loxP: Range, Efficiency, and Somatic Applications. Binhai Zheng, Marijke Sage, Elizabeth A. Sheppeard, Vesna Jurecic, and Allan Bradley. Mol Cell Biol. 2000 Jan; 20(2): 648–655.

GATA3 and the T-cell lineage: essential functions before and after T-helper-2-cell differentiation. I-Cheng Ho, Tzong-Shyuan Tai, and Sung-Yun Pai. Nat Rev Immunol. 2009 Feb; 9(2): 125–135.

Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis. Zhenbang Chen, Lloyd C. Trotman, David Shaffer, Hui-Kuan Lin, Zohar A. Dotan, Masaru Niki, Jason A. Koutcher, Howard I. Scher, Thomas Ludwig, William Gerald, Carlos Cordon-Cardo, and Pier Paolo Pandolfi. Nature. 2005 Aug 4; 436(7051): 725–730.

Terminate and make a loop: regulation of transcriptional directionality. Pawel Grzechnik, Sue Mei Tan-Wong, and Nick J. Proudfoot. Trends Biochem Sci. 2014 Jul; 39(7): 319–327.

Satb1integrates DNA sequence,shape, motif densityand torsional stressto differentially bind targets in nucleosome-dense regions. Rajarshi P. Ghosh, Quanming Shi, Linfeng Yang, Michael P. Reddick, Tatiana Nikitina, Victor B. Zhurkin, Polly Fordyce, Timothy J. Stasevich, Howard Y. Chang, William J. Greenleaf, Jan T. Liphardt. Doi: https://doi.org/10.1101/450262, Biorxiv.

MEME SUITE: tools for motif discovery and searching. Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, *Nucleic Acids Research*, 37:W202-W208, 2009.

Brn-1 and Brn-2 share crucial roles in the production and positioning of mouse neocortical neurons. Sugitani Y, Nakai S, Minowa O, Nishi M, Jishage K, Kawano H, Mori K, Ogawa M, Noda T. Genes Dev. 2002 Jul 15;16(14):1760-5.

Extracellular matrix components of the mouse thymus microenvironment: ontogenetic studies and modulation by glucocorticoid hormones. J Lannes-Vieira, M Dardenne, W Savino. Journal of histochemistry and cytochemistry, Vol 39, Issue 11, 1991.

The thymus microenvironment in regulating thymocyte differentiation. Jacy Gameiro, Patrícia Nagib, and Liana Verinaud. Cell Adh Migr. 2010 Jul-Sep; 4(3): 382–390.

Structural conservation of interferon gamma among vertebrates. Ram Savan, Sarangan Ravichandran, Jack R. Collins, Masahiro Sakai, and Howard A. Young. Cytokine Growth Factor Rev. 2009 Apr; 20(2): 115–124.

Negative feedback loop of Wnt signaling through upregulation of conductin/axin2 in colorectal and liver tumors. Lustig B, Jerchow B, Sachs M, Weiler S, Pietsch T, Karsten U, van de Wetering M, Clevers H, Schlag PM, Birchmeier W, Behrens J. Mol Cell Biol. 2002 Feb;22(4):1184-93.

Identification of c-MYC as a target of the APC pathway. He TC, Sparks AB, Rago C, Hermeking H, Zawel L, da Costa LT, Morin PJ, Vogelstein B, Kinzler KW. Science. 1998 Sep 4;281(5382):1509-12.

PPARdelta is an APC-regulated target of nonsteroidal anti-inflammatory drugs. He TC, Chan TA, Vogelstein B, Kinzler KW. Cell. 1999 Oct 29;99(3):335-45.

Target genes of beta-catenin-T cell-factor/lymphoid-enhancer-factor signaling in human colorectal carcinomas. Mann B, Gelos M, Siedow A, Hanski ML, Gratchev A, Ilyas M, Bodmer WF, Moyer MP, Riecken EO, Buhr HJ, Hanski C. Proc Natl Acad Sci U S A. 1999 Feb 16;96(4):1603-8.

CTLA-4 is a direct target of Wnt/beta-catenin signaling and is expressed in human melanoma tumors. Shah KV, Chien AJ, Yee C, Moon RT. J Invest Dermatol. 2008 Dec;128(12):2870-9. doi: 10.1038/jid.2008.170. Epub 2008 Jun 19.

Wnt/β-catenin signaling: components, mechanisms, and diseases. Bryan T. MacDonald, Keiko Tamai, and Xi He. Dev Cell. 2009 Jul; 17(1): 9–26.

Elevated levels of Wnt signaling disrupt thymus morphogenesis and function. Swann JB, Happe C, Boehm T. Sci Rep. 2017 Apr 11;7(1):785. doi: 10.1038/s41598-017-00842-0.

Kuby Immunology. Kindt, Thomas J, 1939-; Goldsby, Richard A; Osborne, Barbara Anne; Kuby, Janis. Sixth Edition. New York : W.H. Freeman, c2007.

Fast gapped-read alignment with Bowtie 2. Langmead B, Salzberg SL. Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923.

Tsuyuzaki K, Nikaido I (2019). *metaSeq: Meta-analysis of RNA-Seq count data in multiple studies*. R package version 1.24.0.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004 Jan 1;32(Database issue):D493-6.

Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, Jaak Vilo:g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)Nucleic Acids Research 2019;

Zhang et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* (2008) vol. 9 (9) pp. R137

BEDTools: a flexible suite of utilities for comparing genomic features. Aaron R. Quinlan Ira M. Hall. *Bioinformatics*, Volume 26, Issue 6, 15 March 2010, Pages 841–842

Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", *Nucleic Acids Research*, 37:W202-W208, 2009.

Robinson MD, McCarthy DJ, Smyth GK (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” *Bioinformatics*, 26(1), 139-140.

Wei Z, Zhang W, Fang H, Li Y, Wang X (2018). “esATAC: an easy-to-use systematic pipeline for ATAC-seq data analysis.” *Bioinformatics*.

Luo, Weijun, Brouwer, Cory (2013). “Pathview: an R/Bioconductor package for pathway-based data integration and visualization.” *Bioinformatics*, **29**(14), 1830-1831

Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Felix Krueger Simon R. Andrews. *Bioinformatics*, Volume 27, Issue 11, 1 June 2011, Pages 1571–1572.

Feng H, Conneely K, Wu H (2014). “A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data.” Nucleic acids research.

Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. Alicia N. Schep, Jason D. Buenrostro, Sarah K. Denny, Katja Schwartz, Gavin Sherlock, and William J. Greenleaf. Genome Res. 2015 Nov; 25(11): 1757–1770.

Organizational principles of 3D genome architecture. M. Jordan Rowley ,Victor G. Corces. Nature Reviews Genetics 19, 789–800 (2018).

Lineage-determining transcription factor TCF-1 initiates the epigenetic identity of T cell development. John L. Johnson, Georgios Georgakilas, Jelena Petrovic, Makoto Kurachi, Stanley Cai, Christelle Harly, Warren S. Pear, Avinash Bhandoola, E. John Wherry and Golnaz Vahedi. Immunity. 2018 Feb 20; 48(2): 243–257.e10.

TCF-1 and LEF-1 act upstream of Th-POK to promote CD4+ T cell lineage choice and cooperate with Runx3 to silence the Cd4 gene in CD8+ T cells. Farrah C. Steinke, Shuyang Yu, Xinyuan Zhou, Bing He, Wenjing Yang, Bo Zhou, Hiroshi Kawamoto, Jun Zhu, Kai Tan, and Hai-Hui Xue. Nat Immunol. 2014 Jul; 15(7): 646–656.

Aire unleashes stalled RNA polymerase to induce ectopic gene expression in thymic epithelial cells. Matthieu Giraud, Hideyuki Yoshida, Jakub Abramson, Peter B. Rahl, Richard A. Young, Diane Mathis and Christophe Benoist. Proc Natl Acad Sci U S A. 2012 Jan 10; 109(2): 535–540.

Thymic Crosstalk Coordinates Medulla Organization and T-Cell Tolerance Induction. Noëlla Lopes, Arnauld Sergé, Pierre Ferrier and Magali Irla. Front Immunol. 2015; 6: 365

Identification of significant chromatin contacts from HiChIP data by FitHiChIP. Sourya Bhattacharyya, Vivek Chandra, Pandurangan Vijayanand & Ferhat Ay. Nature Communications 10, Article number: 4221 (2019)

Long-range enhancer-promoter contacts in gene expression control. Schoenfelder S, Fraser P. Nat Rev Genet. 2019 Aug;20(8):437-455. doi: 10.1038/s41576-019-0128-0.